CHARACTERISATION OF MOLECULAR NITROGEN IMPLANTED SILICON FOR MULTIPLE THICKNESSES OF GATE OXIDE IN A 0.5μm CMOS PROCESS

Michael Rennie

Doctor of Philosophy in Electrical Engineering

The University of Edinburgh

1996

List of Contents

CHAPTER 1

Introduction

1.1 Background

The key to increased computer processing capability or number of instructions executed per second is to scale the transistor switch to smaller geometries and/or integrate more transistors on a microprocessor integrated circuit. The dominant choice of technology is the silicon Complementary Metal Oxide Semiconductor (CMOS) due to its low power dissipation, high digital functionality, advantage in device scaling, and ease of manufacturability [1]. The speed of a microprocessor CMOS circuit is governed by the transistor drive currents which determine the rate of charge and discharge of a node. Table 1 shows the trend in MOS transistor scaling over the years.

Year of Introduction	1977	1979	1982	1985	1988	1991	1994
Gate Length (μm)	3	2	1.5	1.1	0.9	0.6	0.5
Minimum Feature Size (μm)	3	2	1.5	1	0.7	0.5	0.4
V_{cc} (V)	5	5	5	5	5	5	3.3
MOSFET Gate Oxide (nm)	70	40	25	25	20	13.5	10
Junction Depth (μm)	0.6	0.4	0.3	0.25	0.2	0.15	0.15
NMOS I_{Dsat} @ V_{GS}=5V (mA/μm)	0.1	0.14	0.23	0.27	0.36	0.64	0.82
PMOS I_{Dsat} @ V_{GS}=-5V (mA/μm)		0.06	0.11	0.14	0.19	0.31	0.42
Gate Delay @ FO=1 (ps)	800	350	250	200	160	90	40

Table 1 Evolution of MOSFET technology from 1977 to 1994. Modified from [2].

Scaling CMOS devices down to the sub-half micron channel length increases the device packing density and circuit speed but also introduces many fabrication process problems. Gate oxide thicknesses have also thinned in order to prevent velocity saturation which for gate lengths below 0.5μm, reduces the scaled increase in drive current with reduced gate length [3]. Short channel effects are reduced by scaling the oxide thickness which gives the gate more

control of the channel charge and reducing the shallow junction depth keeps the drain field from extending far into the channel [4]. The thinner gate oxides however are less able to block the diffusion of dopants and impurities and have higher leakage currents compared to thicker dielectrics [5]. Smaller device channel lengths and higher doping concentrations lead to more hot carrier generation. In order to maintain the reliability of MOS transistors the power supply voltage has been lowered at the cost of reduced device performance [6].

Several ways to reduce the hot carrier effects in scaled MOSFET's have been used. These methods include source/drain engineering approaches to reduce the high electric field in the MOS device or move the high electric field away from the vulnerable silicon-silicon oxide interface. Another approach is to harden the oxide from hot carrier damage by the incorporation of nitrogen. These oxides which are often referred to as oxynitride dielectrics, also provide a barrier to the diffusion of impurities into the MOSFET channel. Hardening of the oxide by nitrogen incorporation can be divided into methods where the nitrogen is incorporated during the oxide growth or methods where the nitridation occurs after the oxide is grown.

1.2 Thesis Topic

The aggressive scaling of MOSFET's for speed advantages and reduced power supply voltage to maintain device reliability in microprocessor chips, can result in situations where the memory and other peripheral chips operate at a higher voltage. The data bus from the memory chips can cause severe damage to the thin oxides on the microprocessor chip through Time Dependant Dielectric Breakdown (TDDB) [7].

One solution to overcome this problem is to use circuitry to step down the voltage [8]. A series resistor can be placed at the input to the microprocessor chip to drop the voltage down but a decrease in speed results due to the increased propagation delay. Another method uses an external pull-up resistor in the open-drain output of a CMOS device but this uses up valuable silicon area and can cause excess power dissipation.

Another solution is to produce gate oxides of different thicknesses. The conventional method of accomplishing this involves the masking of the gate oxide with photoresist and etchback of the unmasked gate oxide. This method is susceptible to contamination from the etchback as well as from the resist material. A recently published method [9] has been used to integrate a 16.5nm 5.0V oxide with a 10.0nm 3.3V oxide into a 0.6μm CMOS process by initially growing a 16.5nm oxide then depositing polysilicon and patterning to leave polysilicon on only the 5.0V device areas. The gate oxide is then stripped in the 3.3V device areas and the

thinner 10.0 nm oxide is grown followed by polysilicon deposition and patterning to leave the second poly on only the 10.0nm oxide. Although this process avoids the contact of photoresist with gate oxide, it is very process step intensive and introduces additional thermal steps which can drive the threshold adjust implants which are performed through a sacrificial oxide prior to each gate oxidation.

Recently Soleimani, Doyle and Philipossian [10] have proposed a method of growing two gate oxide thicknesses by the selective implantation of nitrogen through a sacrificial oxide into silicon prior to the sacrificial oxide strip and gate oxidation. In this work, oxides were grown on silicon with different doses of nitrogen implant with and without an implant activation anneal prior to the sacrificial oxide strip. The nitrogen was seen to retard the oxide growth rate depending on nitrogen dose through a build up of nitrogen at the silicon surface. Increased nitrogen dose was also seen to increase the oxide charge which implies increased nitrogen incorporation in the oxide and hence this technique was also proposed to harden gate oxides to hot carrier effects. Samples without the post N-implant anneal were seen to have a higher gate oxide nitrogen content and to retard the oxidation more compared to samples with the anneal.

The topic of research in this thesis is to determine feasibility of integrating the nitrogen implanted silicon without the post N-implant anneal technique into the CMOS5 process. The CMOS5 process is the Digital Equipment Corporation Ltd., 0.5μm drawn channel length 3.3V CMOS process which is used to fabricate the 300MHz Alpha microprocessor and supporting device chip set. The benefits and limitations of this technique from the device performance, reliability and manufacturing perspectives are determined with collaboration from H.R. Soleimani and B.S. Doyle.

1.3 Organization of the Thesis

The growth kinetics and properties of thin gate oxides required for deep submicron CMOS devices are discussed in Chapter 2. This review is intended to serve as a basis for comparing the oxide growth kinetics of molecular nitrogen implanted silicon. Chapter 3 covers MOS theory relevant to the thesis topic. MOS capacitors are discussed with the aid of band theory as a precursor to the explanation of the operation of long channel MOSFET's. The chapter concludes with sections on non-ideal effects and short channel effects which are applicable to the sub half-micron devices used in this work. This provides an understanding of the MOS device operation such that the characteristics of device with the nitrogen implanted silicon technique can be interpreted. The topic of MOS gate oxide yield and reliability is covered in Chapter 4. The affect of plasma processing on the thin oxide reliability through the antenna

effect is also discussed. The reliability of the gate oxide is an essential component of this work that requires a thorough understanding of the fundamental mechanisms.

The phenomena of hot carrier injection and the resulting degradation of MOSFET characteristics with time is discussed in Chapter 5. Some methods to reduce the rate of degradation by trading with the MOSFET performance are included. This chapter also covers methods to modify the gate oxide such that it is hardened to hot carrier effects, including nitrogen incorporation.

The characterisation of the oxide growth kinetics of nitrogen implanted silicon is detailed in Chapter 6 for the experimentation carried out at Digital Equipment Corporation Ltd., Scotland. This is an essential part of this work since the oxide kinetics of nitrogen implanted silicon have not been fully studied for the range of nitrogen doses used here. A mechanism is proposed to describe the oxide growth process. The resulting device performance and reliability characteristics of MOSFET's fabricated for the first time using the oxidation of nitrogen implanted silicon are presented in Chapter 7. The mechanism for the trade-off in MOS device reliability and performance with the differential oxide thickness is investigated in this chapter.

Chapter 8 concludes the thesis by discussing the implementation of the nitrogen implantation process to microprocessor and memory technologies. Suggestions for future work on this topic is also included. An extensive list of references followed by the papers submitted for publication as a result of this work are given in the Appendix.

CHAPTER 2

Thermal Oxidation of Single Crystal Silicon for Thin Gate Oxides

In this chapter the growth kinetics of single crystal silicon thermal oxidation are reviewed. The purpose of this is to lay the foundations on which the growth kinetics of molecular nitrogen implanted silicon can be compared against in Chapter 6. This chapter also covers the terminology of charges associated with silicon oxide growth so that physical mechanisms can be attributed to the resulting electrical deviations of thin oxides used for gate dielectrics in MOS devices.

2.1 Growth Mechanism and Kinetics

2.1.1 Silicon Oxidation

Silica or SiO_2 is found in many forms from crystalline, amorphous to vitreous. The glassy form which is used in silicon IC's [11] exhibits a short range order of atomic structure. This is termed amorphous SiO_2. The structural unit is based around SiO_4.

Silicon dioxide has several uses in the fabrication of silicon semiconductor devices; it is used as an dielectric insulator, a mask against ion implantation or dopant diffusion in silicon, and as a surface passivation layer.

A silicon surface forms into an oxide layer when exposed to an oxidizing ambient due to its high affinity for oxygen. A newly exposed silicon surface will form a very thin (<20Å) 'native' oxide once exposed to air at room temperature.

The most common way of forming silicon oxide is by the thermal oxidation of silicon in oxygen or steam as shown by the chemical reactions in Equations 2.1 and 2.2 respectively.

$$Si(solid) + O_2 \rightarrow SiO_2(solid) \qquad 2.1$$

$$Si(solid) + 2H_2O \rightarrow SiO_2(solid) + 2H_2 \qquad 2.2$$

Thermal oxidation is usually carried out between 700-1300°C. At 960°C oxide becomes viscous and grows stress free.

Other methods such as rapid thermal oxidation, chemical vapor deposition (CVD) and plasma oxidation can also be used. Thermal oxidation however forms the highest quality of oxide albeit with the highest thermal budget.

The chemical reaction to form silicon oxide involves the formation of a silicon-oxygen covalent bond through shared valence electrons. The oxidation of silicon proceeds by the

movement of the Si-SiO_2 interface into the silicon. While it has been established that the oxidizing species diffuse through the oxide to the Si-SiO_2 interface, uncertainty exists as to whether the diffusing species are neutral or charged. The molecular weights and densities of Si and SiO_2 are such that for every thickness x of oxide grown, a thickness of $0.44x$ of silicon is consumed. This means that SiO_2 grows into and out of silicon at roughly the same rate.

2.1.2 Linear-Parabolic Model

The Deal-Grove model [12] describes the kinetics of silicon oxidation for oxide growth >300Å between 700 and 1300°C at partial pressures between 0.2 and 1.0 atmospheres in oxygen or water vapor. The model for the oxidation process is shown in Figure 2.1. The silicon substrate is covered by an oxide layer that is exposed to an oxidizing ambient. The oxidizing species are transported from the gas-phase to the oxide surface with a flux F_1, they then diffuse through the oxide with flux F_2 and then react at the Si-SiO_2 interface with flux F_3. For the steady-state condition used here $F_1 = F_2 = F_3$.

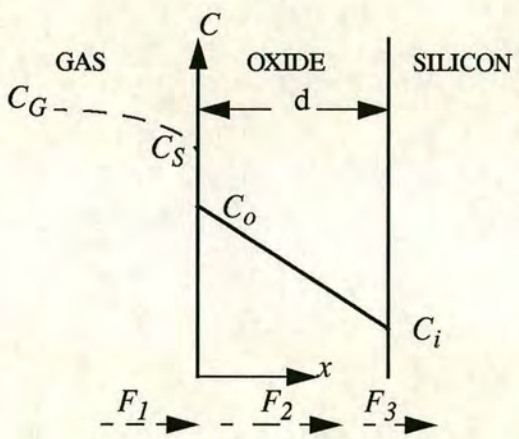


Figure 2.1 Model used for thermal oxidation of silicon [12].

The gas-phase flux F_1 which is the number of molecules or atoms per unit area per unit time, is assumed to be related to the difference in oxidation concentration in the bulk gas C_G and the oxidation concentration adjacent to the oxide surface C_S by

$$F_1 = h_G(C_G - C_s) \qquad 2.3$$

where h_G is the gas-phase mass-transfer coefficient.

Henry's law can be used to relate oxidizing species concentration in the oxide to the

concentration in the gas phase,

$$C_o = Hp_s \qquad 2.4$$

and

$$C^* = Hp_G \qquad 2.5$$

where C_o is the equilibrium concentration at the oxide surface, C^* is the equilibrium bulk concentration in the oxide, p_s is the partial pressure in the gas at the oxide surface, p_G is the partial pressure in the bulk gas and H is Henry's law constant. Using Equations 2.4 and 2.5, F_1 can be expressed as

$$F_1 = h(C^* - C_o) \qquad 2.6$$

where $h = \dfrac{h_G}{HkT}$ and is the gas-phase mass-transfer coefficient in the solid.

F_2 is the flux through the oxide toward the silicon substrate. This flux is determined using Fick's law where a steady-state is assumed to exist within the oxide thickness x_o,

$$F_2 = \frac{D(C_o - C_i)}{x_o} \qquad 2.7$$

where D is the diffusion coefficient and C_i is the oxidizing species concentration in the oxide.

The flux F_3 which describes the Si-SiO_2 reaction rate is proportional to C_i and k_s which is the rate constant for the silicon oxidation chemical reaction.

$$F_3 = k_s C_i \qquad 2.8$$

Once Equations 2.6, 2.7 and 2.8 are equated for the steady-state condition and solved, expressions for C_i and C_o can be found by solving simultaneous equations. Boundary conditions can be applied by considering the case when the diffusitivity is very small or very large. When the diffusitivity of oxidant through the oxide is very small, the limiting factor is supply of oxidant to the interface, then $C_i \rightarrow 0$ and $C_o \rightarrow C^*$. For the case of a very high diffusitivity through the oxide, the reaction of the oxidant with Si limits the rate of SiO_2 formation and $C_i = C_c$.

Combining various equations and incorporating the fact that an oxide layer may initially be

present, the following equation is obtained

$$x_o^2 + Ax_o = B(t+\tau) \qquad 2.9$$

where

$$A = 2D\left[\frac{1}{k_s} + \frac{1}{h}\right] \qquad 2.10$$

$$B = \frac{2DC^*}{N_1} \qquad 2.11$$

$$\tau = \frac{x_i^2 + Ax_i}{B} \qquad 2.12$$

τ represents the shift in time to account for the presence of the initial oxide layer x_i.

Solving for x_o gives

$$x_o = \frac{A}{2}\left\{\left[1 + \frac{4B(t+\tau)}{A^2}\right]^{\frac{1}{2}} - 1\right\} \qquad 2.13$$

For the case of long oxidation times in which the transport of oxidant through the relatively thick oxide x_i is the limiting factor, Equation 2.9 can be reduced to the parabolic relation,

$$x_o = Bt^2 \qquad 2.14$$

B is known as the parabolic rate constant.

For the case of short oxidation times Equation 2.9 is reduced to,

$$x_o = \frac{B}{A}(t+\tau) \qquad 2.15$$

This linear law is the SiO_2 reaction rate limiting mechanism and the term B/A is termed the linear rate constant. Factor which affect the parabolic and linear rate constants include temperature, oxidizing ambient, pressure, silicon dopants and silicon surface orientation. Figure 2.2(a) and 2.2(b) show the temperature dependence of B and B/A respectively for both oxidation in O_2 and H_2O at a constant pressure.

Since the linear and parabolic rate constants are higher for oxidation in steam than in O_2, oxidation proceeds faster in steam. The rate of oxidation is given by

$$\frac{dx_o}{dt} = \frac{B}{2x_o + A} \qquad 2.16$$

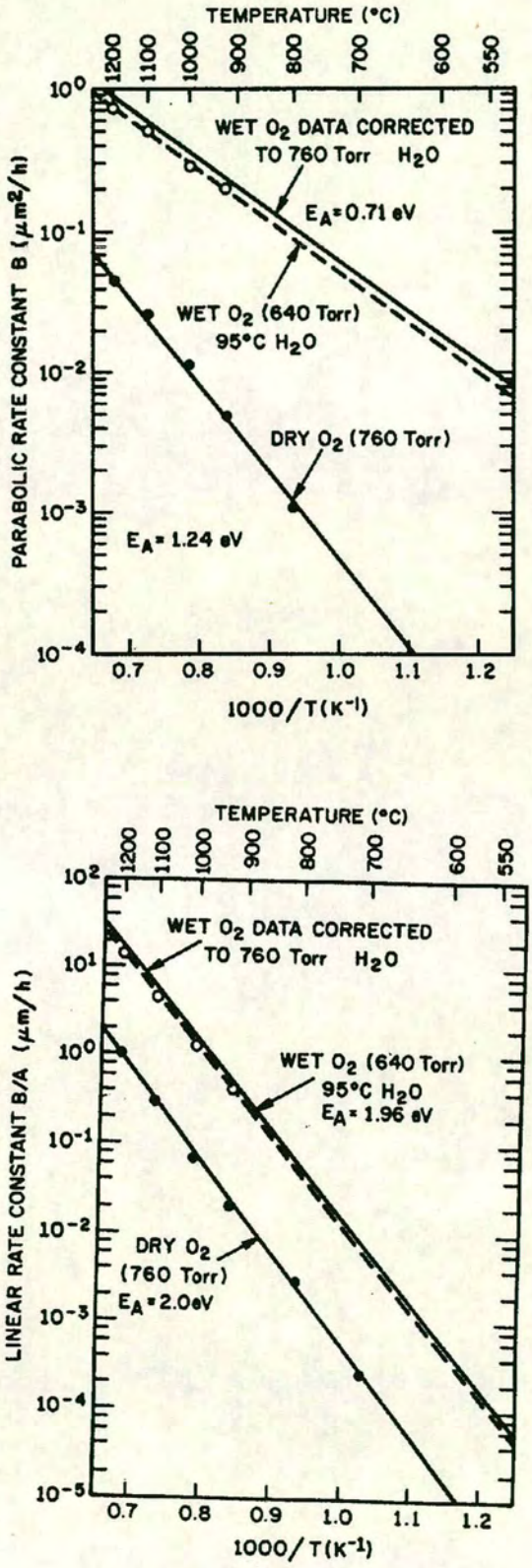


Figure 2.2 The effect of temperature on (a) the parabolic rate constant and (b) the linear rate constant, for dry and wet oxygen. [12]

This model is inadequate to describe oxide growth below 300Å, the model predicts a thickness of ~200Å at zero time. The actual oxidation rate in the thin oxide regime is much higher than that used in the Deal-Grove model. The increased oxidation rate has been speculated to be due to such factors as stress effects, increased oxidant concentration from the high oxidant solubility of bare silicon, high diffusitivity from micropores and field assisted oxidant diffusion.

Massoud et al [13] have included a term to the Deal-Grove model that describes the exponential decay of the oxidation rate with oxide thickness. The modified rate of oxidation is then given by

$$\frac{dx_o}{dt} = \frac{B}{2x_o + A} + C_2 e^{\left(-\frac{x_o}{L_2}\right)} \qquad 2.17$$

The characteristic decay length L_2 is found to be ~70Å and the pre-exponential constant C_2 is 3.2eV for the 800-1000 °C dry oxygen range. The inclusion of this term provides a good fit to thin oxide growth.

An empirical model has been found by Reisman and Nicollian which fits thin oxide growth [14]. This model uses a power law of the form

$$x_o = a(t_g + \tau)^b \qquad 2.18$$

where a and b are constants. τ is the growth time of an existing oxide of thickness x_{oxi} and t_g is the growth time to x_o thickness.

2.1.3 Thin Oxide Growth Methods

MOSFET scaling for performance advantages requires that gate oxides be as thin as 35Å [15]. Oxides as thin as this require to be extremely uniform and reproducible to reduce the variability of device characteristics. This is accomplished by slowing down the oxidation rate While dry oxidation is almost always used for thin oxide growth other conditions are modified.

Lowering the temperature reduces the oxidation rate. However oxides grown at low temperatures have lower quality and higher density than those grown at high temperature. Reduced pressure [16] can also be used to slow down the oxidation rate. Pressures in the range of 0.25 to 2.0 Torr are used and the growth dependence with time is parabolic. High pressure,

low temperature provides a linear growth rate in which the oxide grown is less dense than conventionally grown oxide [17]. The addition of chlorine to the oxidizing ambient can improve device characteristics by reducing the oxide charges. Chlorine can be introduced by various forms such as HCl and TCA (trichloroethane but its corrosive nature limits the concentration used [18].

Silicon wafers are normally inserted into furnace tubes by pulling a quartz boat containing the wafers, in a N_2 ambient at ~700 °C to avoid uncontrolled oxide growth. A few percent of O_2 can be added to the inert ambient to avoid the formation of silicon nitride or micropitting on the Si surface [19]. Once the wafers are loaded into the tube, the temperature is increased to the desired oxidation temperature. The ambient and possibly pressure is then changed to commence oxidation. A post oxidation anneal in an inert ambient is usually performed to reduce the oxide charges. The addition of a few percent O_2 is thought to reduce interface traps. The wafers are cooled down to ~700 °C and subsequently pulled out of the furnace in the inert ambient. Horizontal diffusion furnaces are currently being replaced by vertical furnaces as they provide better temperature uniformity ($<\pm 1$℃) from a more uniform gas flow. Vertical furnaces produce less particles from the wafer loading method used and the fact that dummy wafers can shield other wafers from falling particles [20].

Surface preparation is vitally important to the thin oxide properties [21]. The most common cleaning method is the RCA clean which uses a H_2O-H_2O_2-NH_4OH mixture to remove organic contaminants by the oxidizing nature of the peroxide combined with the solvating action of the ammonium hydroxide. While increasing ammonia concentration can increase the efficiency of contaminant removal, it causes increased silicon microroughness which can degrade the subsequent oxide quality [22]. The technique of Atomic Force Microscopy (AFM) can be used to measure microroughness, this topic will be discussed in more detail later. To remove metallic contaminants, a H_2O-H_2O_2-HCl solution is used where the metals form soluble complexes. An HF dip is usually performed to remove the native oxide layer followed by a rinse in de-ionized water. It has been shown however that an HF-last process provides lower oxide charges and better oxide integrity by passivating the silicon surface with fluorine rather than hydrogen [23]. The method used to terminate the silicon surface has a profound affect on the oxide defect density and device reliability.

Another method [24] to remove the native oxide prior to oxide growth include the use of a cluster tool where the wafers are exposed to an HF vapor prior to being inserted into a furnace under vacuum. The wafers were then moved under vacuum to an low pressure chemical vapor deposition (LPCVD) chamber for polysilicon deposition.

2.2 The Si-SiO_2 interface

The silicon dioxide and SiO_2/Si interface contain fixed and trapped charges as a result of the transition region from amorphous silicon oxide to crystalline silicon. These charges have a profound effect on the MOSFET device characteristics as discussed in Chapter 3 and on the MOSFET device reliability as discussed in Chapters 4 and 5. Not only is it important to reduce the amount of charge associated with the oxidation process, but it is also important to control the variability of the charge levels. The Capacitance-Voltage technique is the most common way of measuring the levels of these charges. This technique will be covered in Chapter 3.

Figure 2.3 shows the four types of charges that exist in the oxide or at the SiO_2/Si interface. The charge Q_{it} is known as the interface trap charge, Q_{ot} is the oxide trapped charge, Q_f is the fixed oxide charge and Q_m is the mobile ionic charge.

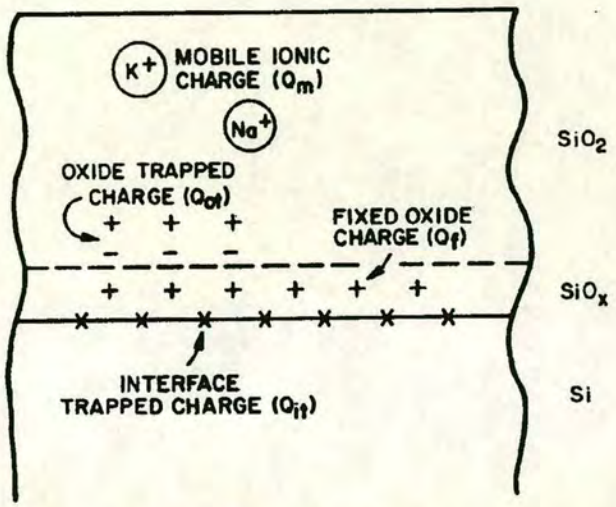


Figure 2.3 Schematic showing the terminology for oxide charges associated with thermally oxidized silicon [25].

2.2.1 Interface Trapped charge

Interface trapped charges are located at the SiO_2-Si interface and have energy states in the forbidden silicon band gap [11]. Each of these energy states is associated with a single silicon atom. The silicon atom is called a trivalent atom and has one dangling bond which can trap free carriers at the silicon surface. Figure 2.4 shows the physical model for interface traps.The magnitude of this charge is a strong function of the oxidizing conditions such as oxidizing ambient and temperature. On [26] the <100> silicon surface there exists 6.8×10^{14} Si atoms per cm^2, if oxidation only left 1/1000 of these dangling bond unsatisfied the density of trapped charge would be 6.8×10^{11} cm^{-2}. If a gate oxide of 200Å was used then the threshold voltage

shift would be 0.63V.

The usual way to reduce the effect of the interface trapped charge is to perform a hydrogen anneal. The dangling bonds are then terminated by a hydrogen bond

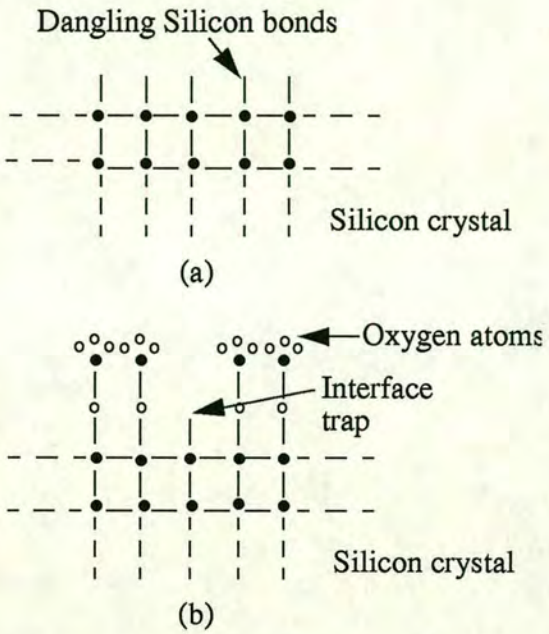


Figure 2.4 Physical model for interface traps.(a) Dangling bonds occur when silicon is abruptly terminated to form a surface. (b) A post oxidation dangling bond that become interface trap. [26]

2.2.2 Fixed Oxide Charge

The fixed oxide charge is located within the oxide in the first 30Å from the SiO_2-Si interface. This charge is usually positive and ranges from 10^{10}-10^{20} cm^{-2} depending on the silicon surface orientation, oxidation process and annealing conditions. The last high temperature step determined the fixed oxide charge density. Inert ambient annealing and rapid cooling from high temperatures results in low values of Q_f [27].

2.2.3 Oxide Trapped Charge

Q_{ot} results from damaged broken bonds and impurities in the silicon dioxide. Damage can also occur from ionized radiation, high current stress through the oxide and avalanche injection. Some sources of ionized radiation that can increase Q_{ot} are sputtering equipment, ion implanters, plasma sources and x-ray photolithography equipment. The charge can be positive or negative and ranges from 10^9-10^{13} cm^{-2}.

2.2.4 Mobile Ion Charge

Mobile alkali ions are the most common source of Q_m. Sodium and potassium can be incorporated into the oxide from handling, dirty quartzware and cleaning chemical impurities. These ions drift through the oxide with applied biases and result in unstable device characteristics. The C-V bias-temperature method is used to measure the mobile ion content. Modern equipment and practices has resulted in significantly low levels of mobile ions, levels of 10^{10} cm^{-2} are considered low. The addition of Cl to the oxidizing ambient can neutralize sodium ions and result in stable devices.

CHAPTER 3

MOS Device Theory

This chapter is devoted to describing the operation of the MOS transistor and the effect of device scaling to dimensions which are relevant to this work. This understanding of the MOS device characteristics is required to interpret the results of the experimentation given in Chapter 7.

3.1 The Metal-Oxide-Semiconductor (MOS) Capacitor [28]

The heart of the MOSFET is the metal-oxide-semiconductor capacitor shown in Figure 3.1. The metal may be sputtered aluminium or some other type of metal although in most cases, it is a high-conductivity (highly-doped) polycrystalline silicon that has been deposited on the oxide using CVD. The parameter t_{ox} in the figure is the thickness of the silicon oxide and ε_{ox} is the permittivity of the oxide.

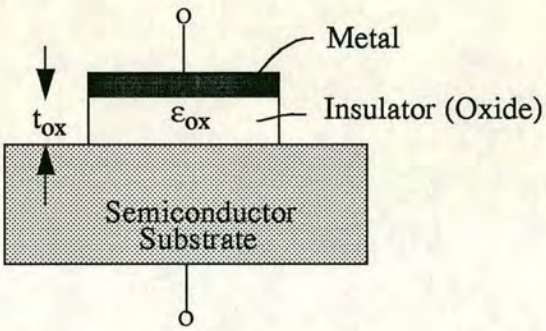


Figure 3.1 The basic MOS capacitor structure.

3.1.1 Energy-Band Diagrams [29]

The physics of the MOS structure can be easily explained with the aid of Figure 3.2a which illustrates the case of the p-type semiconductor substrate. The top metal gate is at a negative voltage with respect to the semiconductor substrate. The majority holes in the semiconductor experience a force toward the semiconductor-oxide interface and an electric field is induced across the oxide. The equilibrium distribution of charge in the MOS capacitor with this particular applied voltage is shown in Figure 3.2b. The accumulation of holes corresponds to the positive charge on the bottom 'plate' of the MOS capacitor.

Figure 3.3a shows the same MOS capacitor in which the polarity of the applied voltage is reversed. A positive charge now exists on the top metal plate and the induced electric field is in the opposite direction as shown. If the electric field penetrates the semiconductor in this case, majority carrier holes will be forced away from the semiconductor-oxide interface and a negative space charge region is created due to the fixed ionized acceptor atoms. The negative

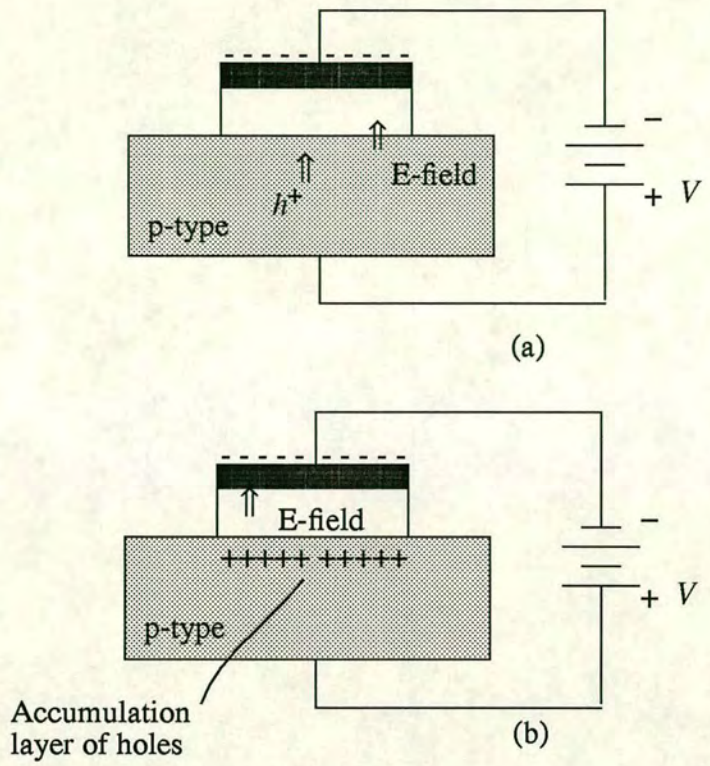


Figure 3.2 (a) The MOS capacitor with a negative gate bias showing the electric field and charge flow. (b) The MOS capacitor with an accumulation layer of holes.

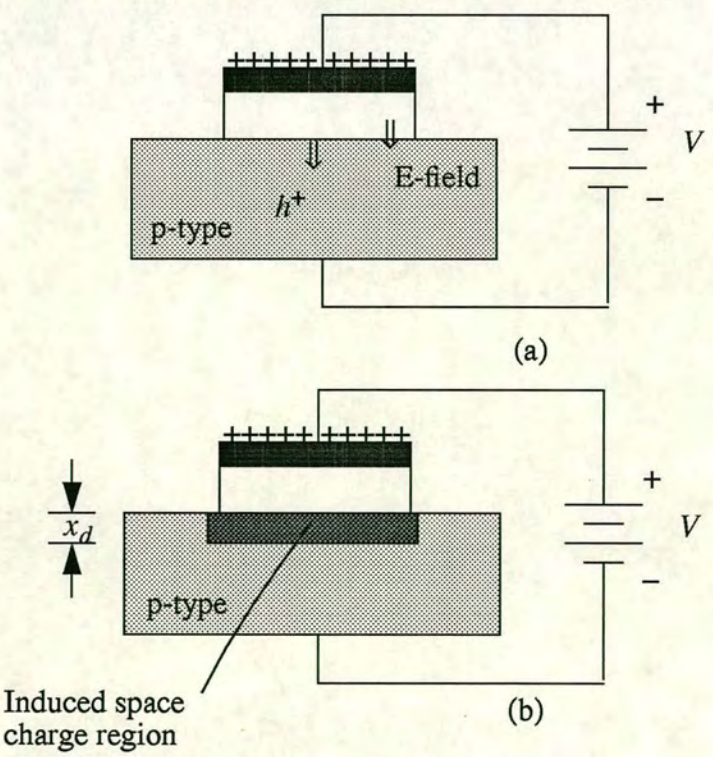


Figure 3.3 The MOS capacitor with a moderate positive gate bias, showing (a) the electric field and charge flow, and (b) the induced space charge region.

charge in the induced depletion region corresponds to the negative charge on the bottom 'plate' of the MOS capacitor. Figure 3.3b shows the equilibrium distribution of charge in the MOS capacitor with this applied voltage.

The energy-band diagram of the MOS capacitor with the p-type substrate, for the case when a negative voltage is applied to the top metal gate, is shown in Figure 3.4a. The valence-band edge is closer to the Fermi level at the oxide-semiconductor interface than in the bulk material, which implies that there is an accumulation of holes. The semiconductor surface appears to be more p-type than the bulk material. The Fermi level is a constant in the semiconductor since the MOS system is in thermal equilibrium and there is no current through the oxide.

Figure 3.4b shows the energy-band diagram of the MOS system when a positive voltage is applied to the gate. The conduction and valence band edges bend as shown in the figure, indicating a space charge region similar to that in a pn junction. The conduction band and intrinsic Fermi levels move closer to the Fermi level. The induced space charge width is given by x_d.

Figure 3.4c shows the condition when a larger positive charge is applied across the MOS capacitor. More band bending occurs and a larger induced space charge region is formed. The intrinsic Fermi level at the surface is now below the Femi level, thus the conduction band is closer to the Fermi level than the valence band is. This result implies that the surface in the semiconductor adjacent to the oxide-semiconductor interface is n-type. By applying a sufficiently large positive gate voltage, the surface of the semiconductor has inverted from p-type to n-type. An inversion layer of electrons has been created at the oxide-semiconductor interface.

In a similar way the band diagrams can be constructed for the case of an n-type substrate. When a positive voltage is applied to the top of the metal gate, an accumulation of electrons is formed at the oxide-semiconductor interface. A negative voltage applied to the surface of the metal gate depletes the oxide-semiconductor interface of majority electrons and a positive space charge region is formed. Increasing the negative voltage further bends the energy bands further and an inversion layer of holes is induced at the oxide-semiconductor interface.

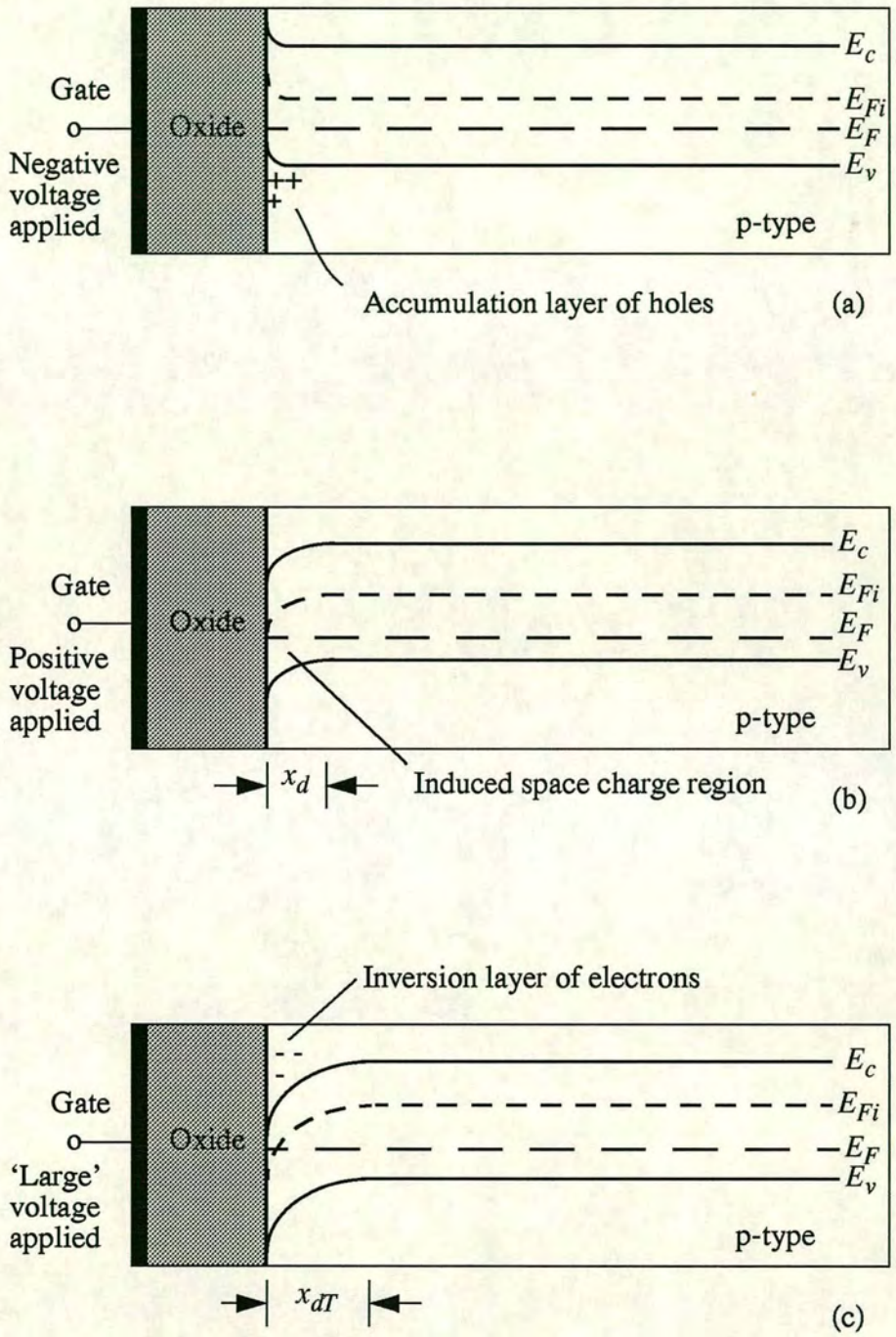


Figure 3.4 The energy-band diagram of a MOS capacitor with a p-type substrate for (a) a negative gate bias, (b) a moderate positive gate bias and (c) a 'large' positive gate bias.

3.1.2 Depletion Layer Thickness

The width of the induced space charge region x_d is shown again in Figure 3.5

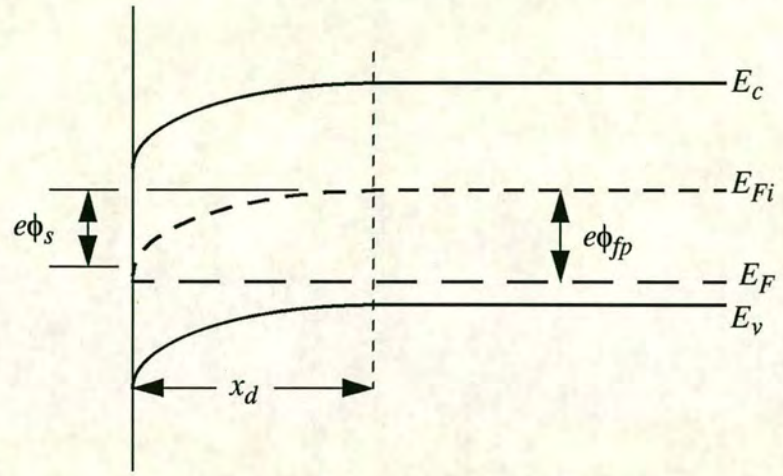


Figure 3.5 The energy-band diagram in the p-type semiconductor, indicating surface potential.

The potential ϕ_{fp} is the difference (in volts) between E_{Fi} and E_F and is given by

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) \qquad 3.1$$

where N_a is the acceptor doping concentration and n_i is the intrinsic carrier concentration.

The surface potential ϕ_s is the difference (in volts) between E_{Fi} measured in the bulk semiconductor and E_{Fi} measured at the surface. The surface potential is the potential difference across the space charge layer. The space charge width is expressed below from that of a one-sided pn junction, assuming that the abrupt depletion region approximation is valid

$$x_d = \left\{\frac{2\varepsilon_s \phi_s}{eN_a}\right\}^{1/2} \qquad 3.2$$

where ε_s is the permittivity of the semiconductor.

Figure 3.6 shows the energy bands for the case in which $\phi_s = 2\phi_{fp}$.

The Fermi level at the surface is as far above the intrinsic level as the Fermi level is below the intrinsic level in the bulk semiconductor. The electron concentration at the surface is the same as the hole concentration in the bulk material. This condition is known as the *threshold*

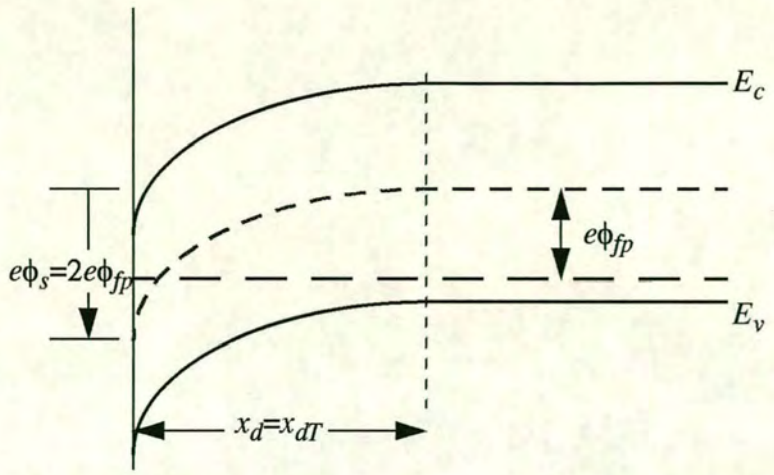


Figure 3.6 The energy-band diagram in the p-type semiconductor at the threshold inversion point.

inversion point. The applied gate voltage creating this condition is known as the *threshold voltage.* If the gate voltage increases above this threshold value, the conduction band will bend slightly closer to the Fermi level, but the change in the conduction band at the surface is now only a slight function of gate voltage. The electron concentration at the surface, however, is an exponential function of the surface potential. The surface potential may increase by a few (kT/e) volts, which will change the electron concentration by orders of magnitude, but the space charge width changes only slightly. In this case, then, the space charge region has essentially reached a maximum width.

The maximum space charge width, x_{dT}, at inversion transition point can be calculated from Equation 3.2 by setting $\phi_s = 2\phi_{fp}$. Then

$$x_{dT} = \left\{\frac{4\varepsilon_s\phi_{fp}}{eN_a}\right\}^{1/2} \qquad 3.3$$

We have been considering a p-type semiconductor substrate. The same maximum induced space charge region width occurs in an n-type substrate. Figure 3.7 is the energy-band diagram at the threshold voltage with an n-type substrate.

Thus

$$\phi_{fn} = V_t \ln\left(\frac{N_d}{n_i}\right) \qquad 3.4$$

and

$$x_{dT} = \left\{ \frac{4\varepsilon_s \phi_{fn}}{eN_d} \right\}^{1/2} \qquad 3.5$$

Figure 3.8 is a plot of x_{dT} at $T = 300^{\circ}$K as a function of doping concentration in silicon. The semiconductor doping can be either n-type or p-type.

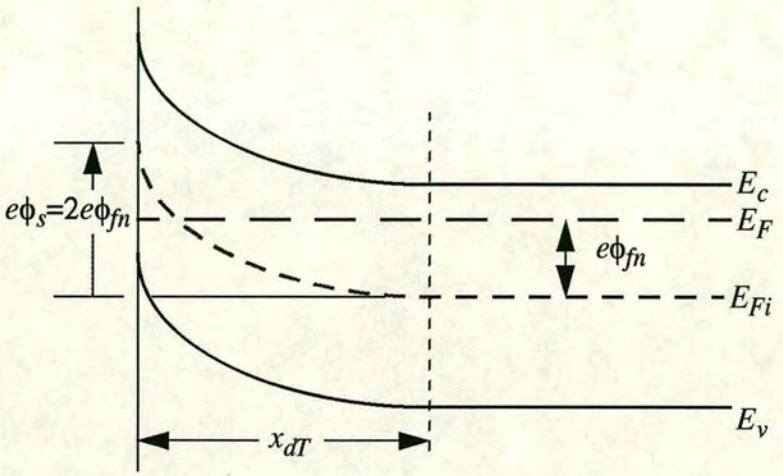


Figure 3.7 The energy-band diagram in the n-type semiconductor at the threshold inversion point.

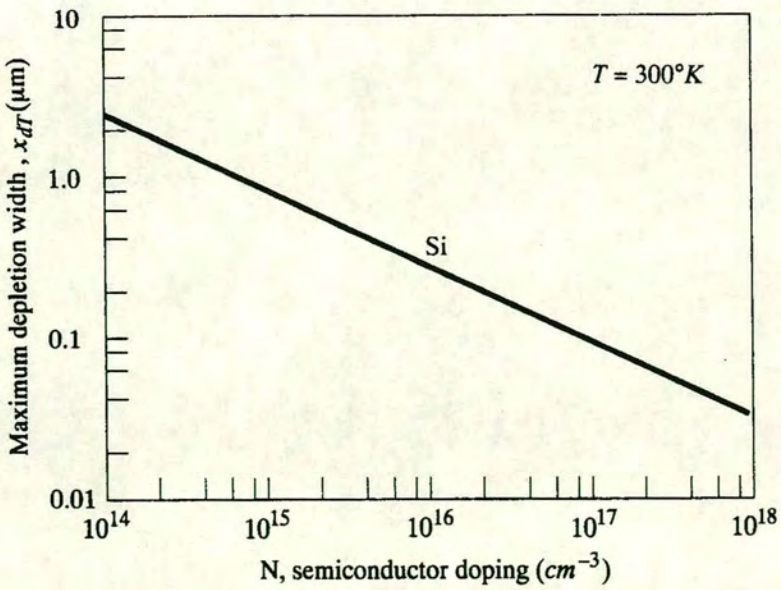


Figure 3.8 Maximum induced space charge region width verses semiconductor doping [30].

3.1.3 Work Function Differences

The energy levels are shown in the metal, silicon dioxide, and silicon relative to the vacuum level in Figure 3.9a. The metal work function is ϕ_m and the electron affinity is χ. The parameter χ_i is the oxide electron affinity and, for silicon dioxide, $\chi_i = 0.9$ volts.

Figure 3.9b shows the energy-band diagram of the entire metal-oxide-semiconductor structure with zero gate voltage applied. The Fermi level is a constant through the entire system at thermal equilibrium. The modified metal work function, ϕ_m', is defined as the potential required to inject an electron from the metal into the conduction band of the oxide. Similarly, χ' is defined as a modified electron affinity. The voltage V_{ox0} is the potential drop across the oxide for zero applied gate voltage and is not necessarily zero because of the difference between ϕ_m and χ. The potential ϕ_{s0} is the surface potential for this case.

The sum of the energies from the Fermi level on the metal side to the Fermi level on the semiconductor side is,

$$e\phi_m' + eV_{ox0} = e\chi' + \frac{E_g}{2} - e\phi_{s0} + e\phi_{fp} \qquad 3.6$$

This equation can be re-written as

$$V_{ox0} + \phi_{s0} = -\left[\phi_m' - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] \qquad 3.7$$

The potential ϕ_{ms} is defined as,

$$\phi_{ms} \equiv \left[\phi_m' - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] \qquad 3.8$$

which is called the metal-semiconductor work function difference.

Degenerately doped polysilicon deposited on the oxide is usually used in place of a metal gate. Figure 3.10a shows the energy-band diagram of a MOS capacitor with an n^+ polysilicon gate and a p-type substrate. Figure 3.10b shows the energy-band diagram for the case of a p^+ polysilicon gate and the p-type silicon substrate. In the degenerately doped polysilicon we assume that, $E_F = E_c$ for the n^+ case and $E_F = E_v$ for the p^+ case.

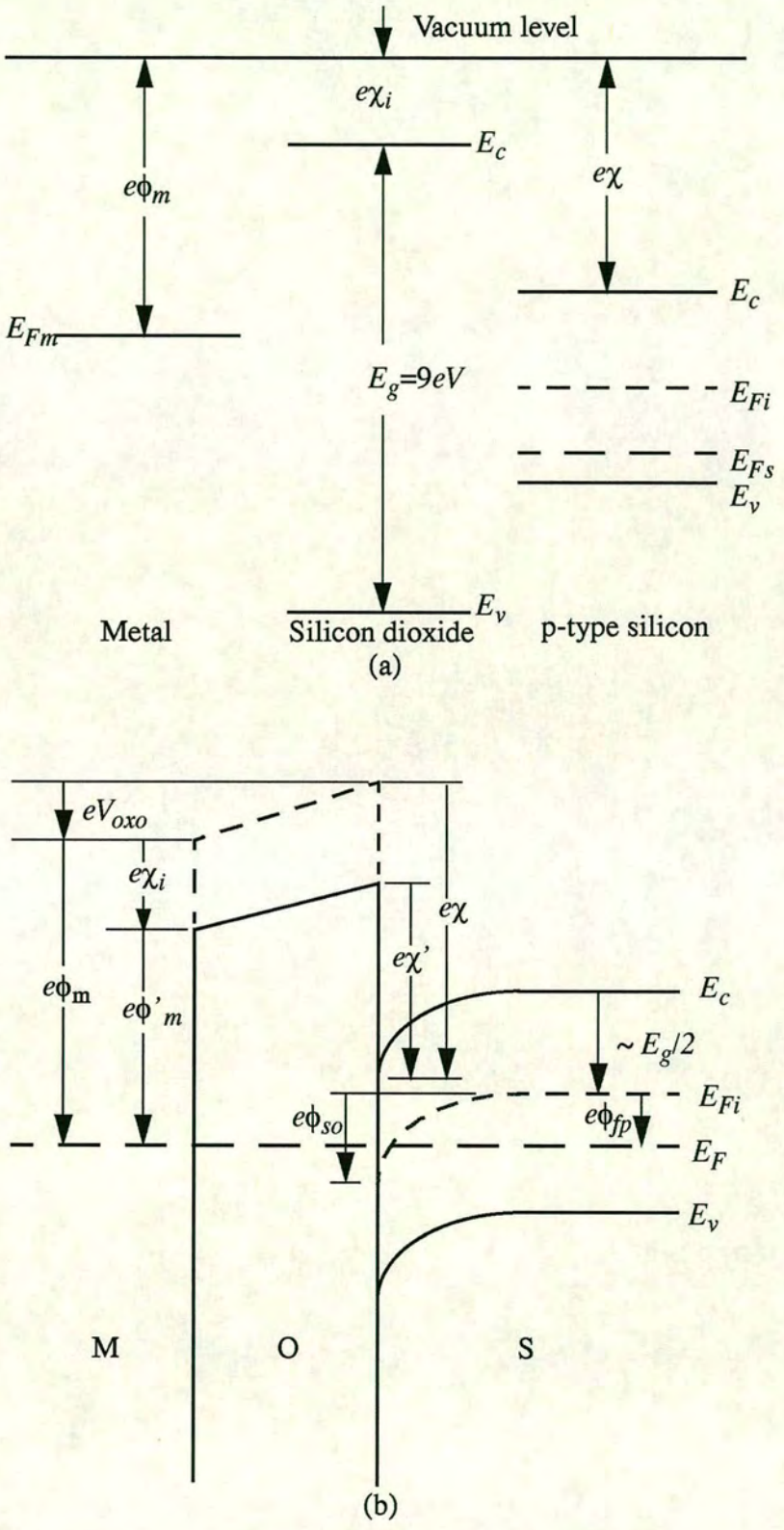


Figure 3.9 (a) Energy levels in a MOS system prior to contact. (b) Energy-band diagram through the MOS structure in thermal equilibrium after contact.

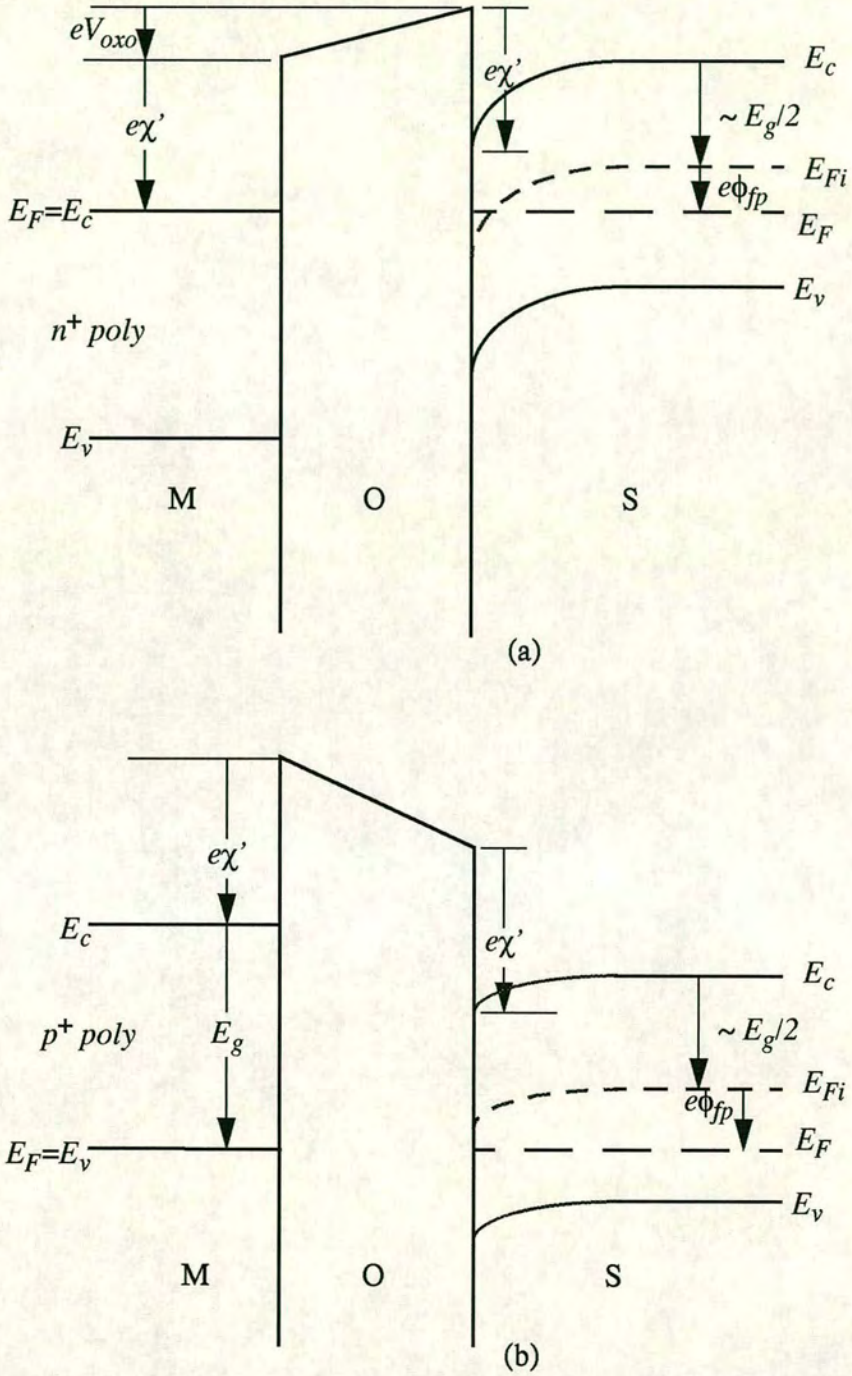


Figure 3.10 Energy-band diagram through the MOS structure with a p-type substrate at zero gate bias for (a) an n^+ polysilicon gate and (b) a p^+ polysilicon gate.

For the n^+ polysilicon gate, the metal-semiconductor work function difference is,

$$\phi_{ms} = \left[\chi' - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] = -\left(\frac{E_g}{2e} + \phi_{fp}\right) \qquad 3.9$$

and for the p^+ polysilicon gate,

$$\phi_{ms} = \left[\left(\chi' + \frac{E_g}{e}\right) - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] = \left(\frac{E_g}{2e} + \phi_{fp}\right) \qquad 3.10$$

The Fermi level however can be 0.1 to 0.2 volt above E_c or below E_v, and so the experimentally obtained values for ϕ_{ms} can be different.

We can also consider the situation where the substrate is a n-type semiconductor and the gate is metal, Figure 3.11. The metal-semiconductor work function difference is defined as,

$$\phi_{ms} \equiv \left[\phi_m' - \left(\chi' + \frac{E_g}{2e} - \phi_{fn}\right)\right] \qquad 3.11$$

where ϕ_{fn} is assumed to be a positive value. Similar expressions can be obtained for n^+ and p^+ polysilicon gates.

Figure 3.12 shows the work function differences as a function of semiconductor doping for the various types of gates. The magnitude of ϕ_{ms} for the polysilicon gates is larger than what Equation 3.9 and Equation 3.10 predict. This difference is due to the Fermi level not being equal to the conduction band energy for the n^+ gate or the valence band energy for the p^+ gate.

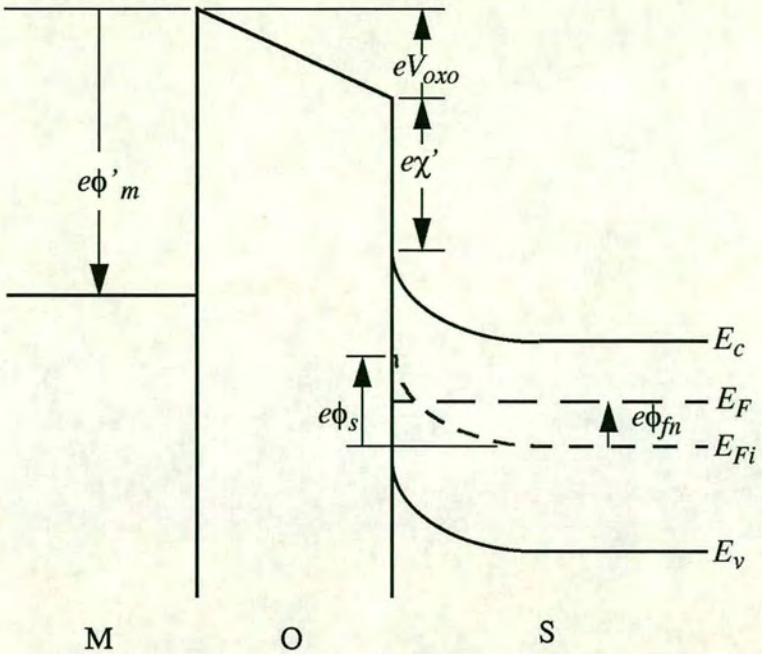


Figure 3.11 Energy-band diagram through the MOS structure with an n-type substrate for a negative applied gate bias.

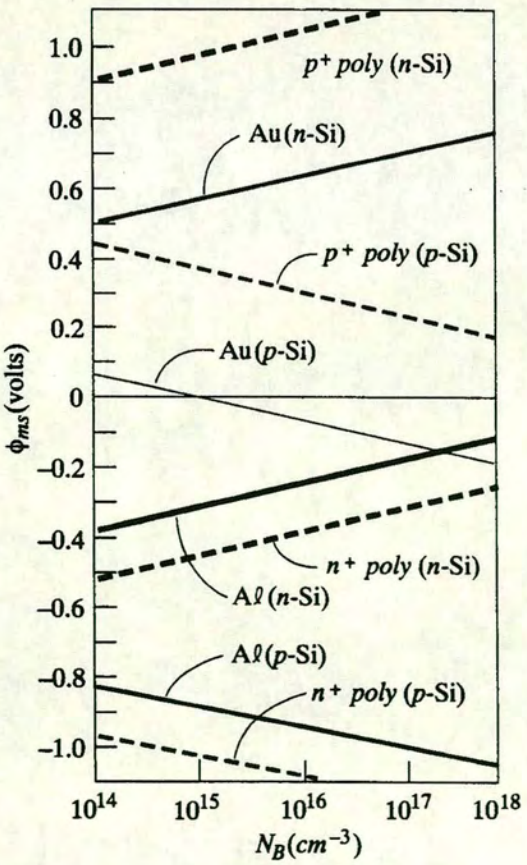


Figure 3.12 Metal-semiconductor work function difference verses doping for aluminium, gold, n^+-, and p^+-polysilicon gates [30].

3.1.4 Flat-Band Voltage

The flat-band voltage is defined as the applied gate voltage such that there is no band bending in the semiconductor and there is no space charge region. Figure 3.13 shows the flat-band condition. The voltage across the oxide is not necessarily zero because of the work function difference and possible trapped charge in the oxide as discussed in Chapter 1.

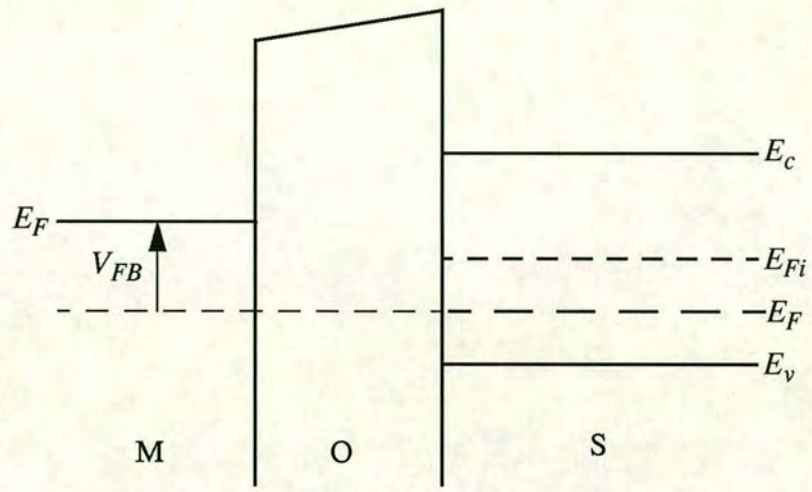


Figure 3.13 Energy-band diagram of a MOS capacitor at flat-band.

The net fixed charge in the oxide appears to be located fairly close to the oxide-semiconductor interface. In the following analysis of the MOS structure we shall assume that an equivalent trapped charge per unit area, Q_{ss}', is located in the oxide directly adjacent to the oxide-semiconductor interface. Other oxide-type charges that may exist in the device will be ignored for now.

Equation 3.7, for a zero applied gate voltage, can be written as,

$$V_{ox0} + \phi_{s0} = -\phi_{ms} \qquad 3.12$$

If a gate voltage is applied, the potential drop across the oxide and the surface potential will change as,

$$V_G = \Delta V_{ox} + \Delta\phi_s = (V_{ox} - V_{ox0}) + (\phi_s - \phi_{s0}) \qquad 3.13$$

Using Equation 3.12,

$$V_G = V_{ox} + \phi_s + \phi_{ms} \qquad 3.14$$

Figure 3.14 shows the charge distribution in the MOS structure for the flat-band condition. There is zero net charge in the semiconductor and an equivalent fixed surface charge density exists in the oxide. The charge density on the metal is Q_m' and from charge neutrality,

$$Q_m' + Q_{ss}' = 0 \qquad 3.15$$

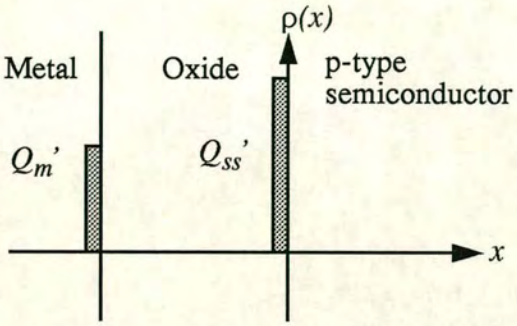


Figure 3.14 Charge distribution in a MOS capacitor at flat-band.

Q_m' can be related to the voltage across the oxide by,

$$V_{ox} = \frac{Q_m'}{C_{ox}} \qquad 3.16$$

where C_{ox}' is the oxide capacitance per unit area. Substituting Equation 3.15 into Equation 3.16

$$V_{ox} = \frac{-Q_{ss}'}{C_{ox}} \qquad 3.17$$

In the flat-band condition, the surface potential is zero, or $\phi_s = 0$. Then from Equation 3.14, the flat-band voltage for the MOS device is,

$$V_G = V_{FB} = \phi_{ms} - \frac{Q_{ss}'}{C_{ox}} \qquad 3.18$$

3.1.5 Threshold Voltage

The threshold voltage was defined as the applied gate voltage required to achieve the threshold inversion point. The threshold inversion point is defined as the condition when the surface potential is $\phi_s = 2\phi_{fp}$ for the p-type semiconductor and $\phi_s = 2\phi_{fn}$ for the n-type semiconductor.

Figure 3.15 shows the charge distribution through the MOS device at the threshold inversion point for a p-type semiconductor substrate. The space charge width has reached its maximum value. There is an equivalent oxide charge Q_{ss}' and the positive charge on the metal gate at threshold is Q_{mT}'. From conservation of charge,

$$Q_{mT}' + Q_{ss}' = |Q_{SD}'(max)| \qquad 3.19$$

where

$$|Q_{SD}'(max)| = eN_a x_{dT} \qquad 3.20$$

and is the magnitude of the maximum space charge density per unit area of depletion region.

The energy-band diagram of the MOS system with an applied positive gate voltage is shown in Figure 3.16. The applied gate voltage will change the voltage across the oxide and will change the surface potential. At threshold, $V_G = V_{TN}$ is defined as the threshold voltage that creates the electron inversion layer charge. The surface potential is $\phi_s = 2\phi_{fp}$ at threshold and equation 3.13 can be written as,

$$V_{TN} = V_{oxT} + 2\phi_{fp} + \phi_{ms} \qquad 3.21$$

where V_{oxT} is the voltage across the oxide at this threshold inversion point. The voltage V_{oxT} can be related to the charge on the metal and to the oxide capacitance by,

$$V_{oxT} = \frac{Q_{mT}'}{C_{ox}} \qquad 3.22$$

where C_{ox} is the oxide capacitance per unit area. Using Equation 3.19,

$$V_{oxT} = \frac{Q_{mT}'}{C_{ox}} = \frac{1}{C_{ox}}(|Q_{SD}'(max)| - Q_{ss}') \qquad 3.23$$

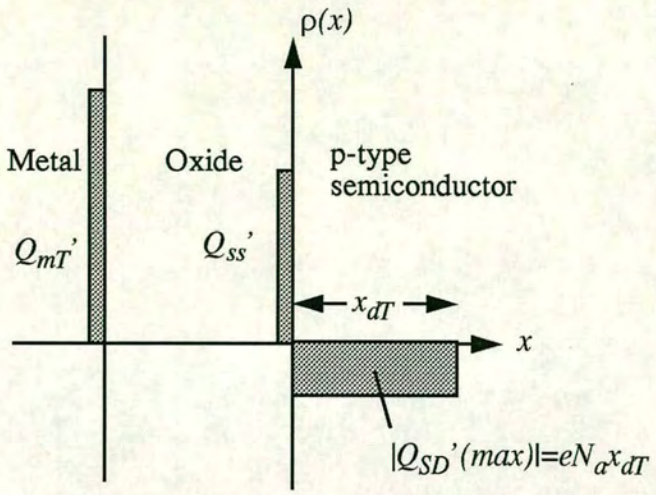


Figure 3.15 Charge distribution in a MOS capacitor with a p-type substrate at the threshold inversion point.

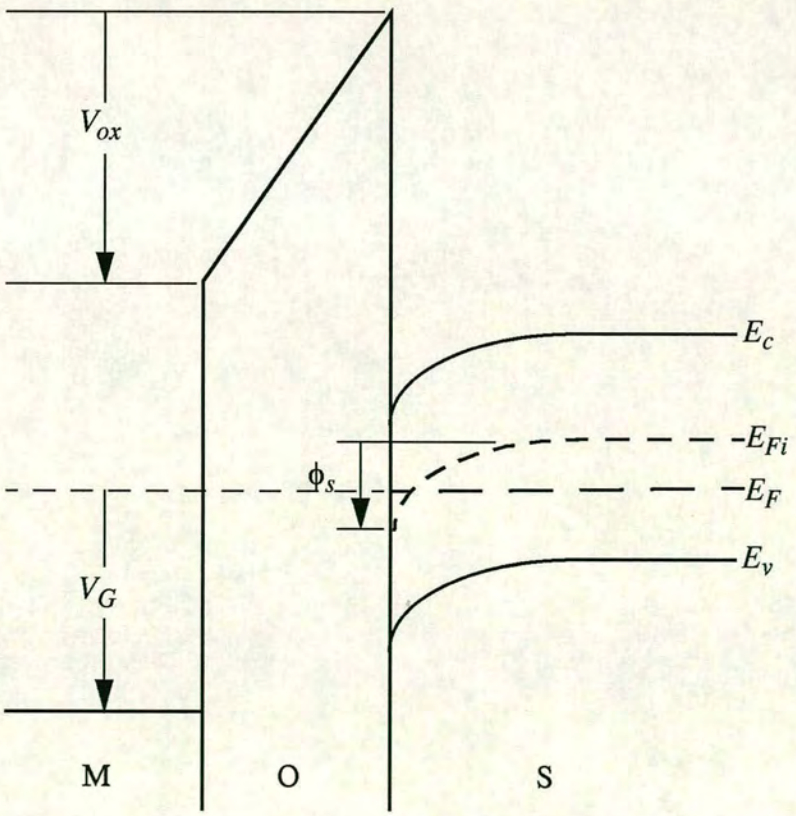


Figure 3.16 Energy-band diagram through the MOS structure with a positive applied gate bias.

The threshold voltage can now be written as

$$V_{TN} = (|Q_{SD}'(max)| - Q_{ss}')\left(\frac{t_{ox}}{\varepsilon_{ox}}\right) + \phi_{ms} + 2\phi_{fp} \qquad 3.24$$

The threshold voltage is the point at which the transistor turns on. If a circuit is to operate between 0 and 5 volts and the threshold of a MOSFET is say 7 volts, the transistor can not be turned on and off and would not act as switch.

A negative threshold voltage for a p-type substrate implies a depletion mode device in which the device can never be turned off for positive applied gate voltages.

Figure 3.17 is a plot of threshold voltage V_{TN} as a function of the acceptor doping concentration for various positive oxide charge values. The p-type semiconductor must be heavily doped to act as an enhancement mode device.

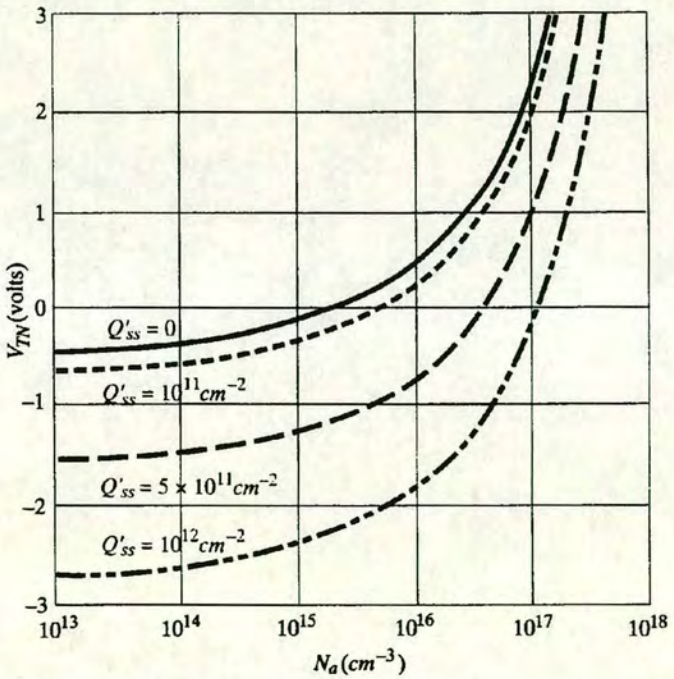


Figure 3.17 Threshold voltage of an n-channel MOSFET verses the p-type substrate doping concentration for various values of oxide trapped charge (t_{ox}=500Å, Al gate)[30].

The same type of derivation can be done with a n-type semiconductor substrate, where a negative gate voltage can induce an inversion layer of holes at the oxide-semiconductor interface. The energy-band diagram for the MOS structure with a n-type substrate and negative applied gate voltage was shown in Figure 3.11. The threshold voltage for this case

can be derived as

$$V_{TP} = (-|Q_{SD}'(max)| - Q_{ss}')\left(\frac{t_{ox}}{\varepsilon_{ox}}\right) + \phi_{ms} - 2\phi_{fn} \qquad 3.25$$

where

$$\phi_{ms} \equiv \left[\phi_m' - \left(\chi' + \frac{E_g}{2e} - \phi_{fn}\right)\right] \qquad 3.11$$

$$|Q_{SD}'(max)| = eN_d x_{dT} \qquad 3.26$$

$$x_{dT} = \left\{\frac{4\varepsilon_s \phi_{fn}}{eN_d}\right\}^{1/2} \qquad 3.5$$

and

$$\phi_{fn} = V_t \ln\left(\frac{N_d}{n_i}\right) \qquad 3.4$$

The quantities x_{dT} and ϕ_{fn} are positive.

Figure 3.18 is a plot of V_{TP} verses doping concentration for several values of Q_{ss}'. As Q_{ss}' increases, the threshold voltage becomes more negative and it takes a larger applied gate voltage to create the inversion layer of holes at the oxide-semiconductor interface.

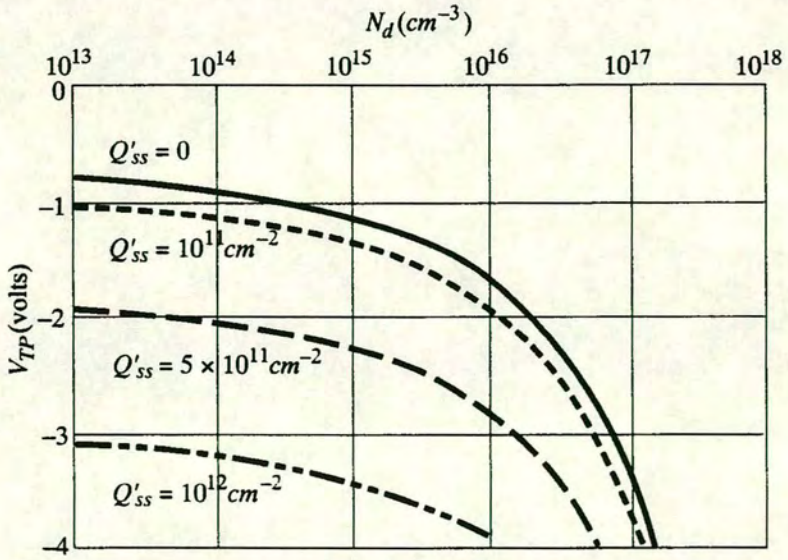


Figure 3.18 Threshold voltage of an p-channel MOSFET verses the n-type substrate doping concentration for various values of oxide trapped charge (t_{ox}=500Å, Al gate)[30].

3.2 Capacitance-Voltage Characteristics [28]

The MOS capacitor structure is the heart of the MOSFET. Lots of information can be obtained about the MOS device and the oxide-semiconductor interface from the capacitance verses voltage or C-V characteristics of the device. The capacitance of a device is defined as,

$$C = \frac{dQ}{dV} \qquad 3.27$$

where dQ is the magnitude of the differential change in charge on one plate as a function of the differential change in voltage dV across the gate capacitor. The capacitance is a small-signal or A.C. parameter and is measured by super-imposing a small ac voltage on an applied D.C. gate voltage. The capacitance, then, is measured as a function of the applied D.C. gate voltage.

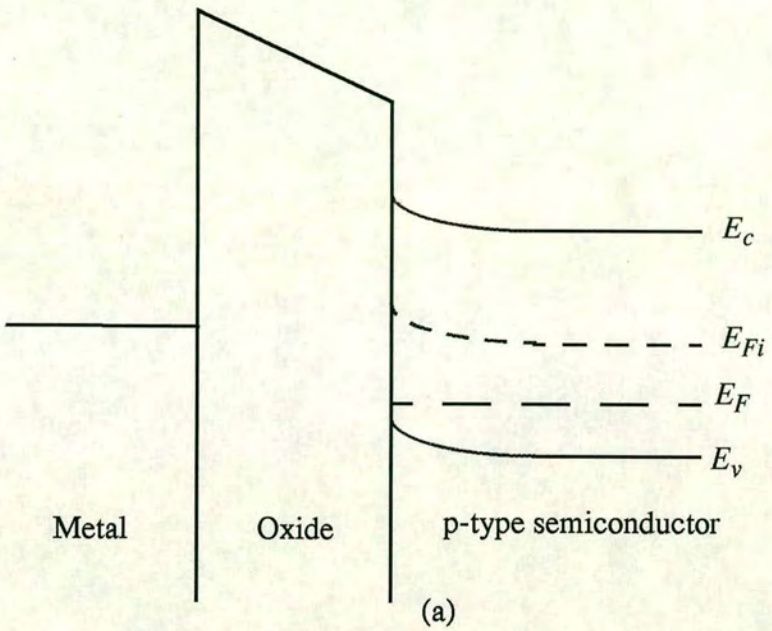
3.2.1 Ideal C-V Characteristics

In the following consideration of the ideal C-V characteristics of the MOS capacitor, there is assumed to be zero charge trapped in the oxide and also no charge trapped at the oxide-semiconductor interface.

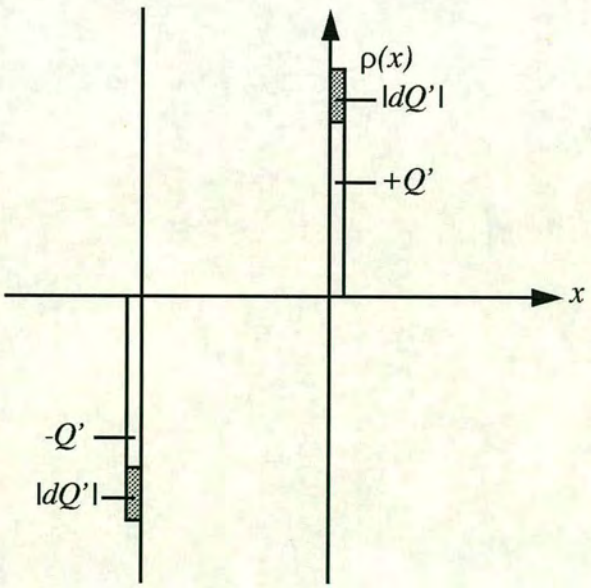
There are three operating conditions of interest in the MOS capacitor; accumulation, depletion and inversion. Figure 3.19a shows the energy-band diagram of a MOS capacitor with a p-type substrate for the case when a negative voltage is applied to the gate, inducing an accumulation layer of holes in the semiconductor at the oxide-semiconductor interface. A small differential change in voltage across the MOS structure will cause a differential change in charge on the metal gate and also in the hole accumulation charge, as shown in Figure 3.19b. The differential changes in charge density occur at the edges of the oxide, as in a parallel-plate capacitor. The capacitance C' per unit area of the MOS capacitor for this accumulation mode is the oxide capacitance,

$$C'(acc) = C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \qquad 3.28$$

Figure 3.20a shows the energy-band diagram of the MOS device when a small positive voltage is applied to the gate inducing a space charge region in the semiconductor. Figure 3.20b shows the charge distribution through the device for this condition.



(a)



(b)

Figure 3.19 (a) Energy-band diagram through a MOS capacitor for the accumulation mode. (b) Differential charge distribution at accumulation for a differential change in gate voltage.

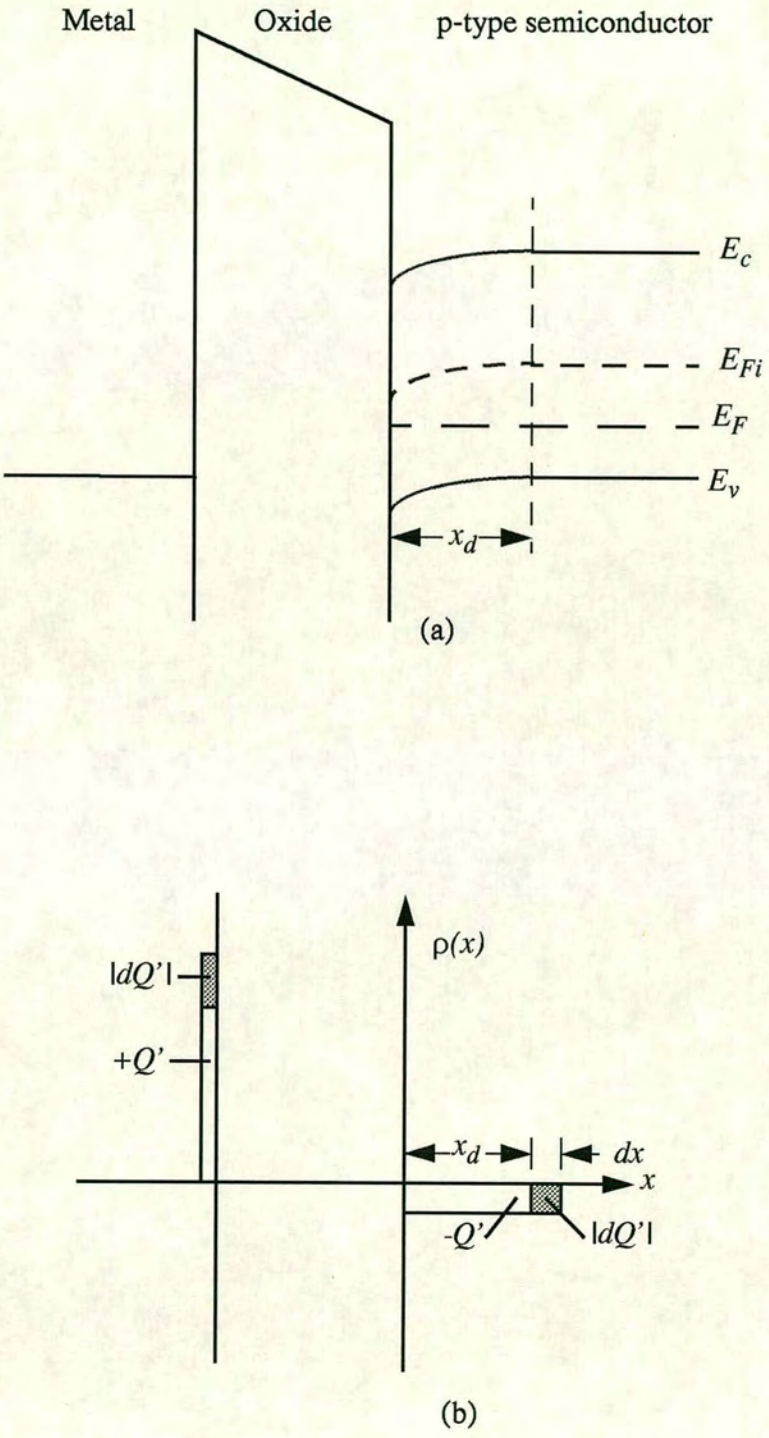


Figure 3.20 (a) Energy-band diagram through a MOS capacitor for the depletion mode. (b) Differential charge distribution at depletion for a differential change in gate voltage.

The oxide capacitance and the capacitance of the depletion region are in series and so the total capacitance is

$$\frac{1}{C'(depl)} = \frac{1}{C_{ox}} + \frac{1}{C_{SD}'} \qquad 3.29$$

or

$$C'(depl) = \frac{C_{ox}C_{SD}'}{C_{ox} + C_{SD}'} \qquad 3.30$$

Since $C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$ and $C_{SD}' = \frac{\varepsilon_s}{x_d}$. Then

$$C'(depl) = \frac{C_{ox}}{1 + \frac{C_{ox}}{C_{SD}}} = \frac{\varepsilon_{ox}}{t_{ox} + \left(\frac{\varepsilon_{ox}}{\varepsilon_s}\right)x_d} \qquad 3.31$$

As the space charge width increases, the total capacitance *C'(depl)* decreases. At the point where the maximum depletion width is reached, the capacitance will be a minimum,

$$C_{min}' = \frac{\varepsilon_{ox}}{t_{ox} + \left(\frac{\varepsilon_{ox}}{\varepsilon_s}\right)x_{dT}} \qquad 3.32$$

Figure 3.21a shows the energy-band diagram of this MOS device for the inversion condition. In the ideal case, a small incremental change in the voltage across the MOS capacitor will cause a differential change in the inversion layer density. The space charge width does not change. If the inversion charge can respond to the change in capacitance voltage as indicated in Figure 3.21b, then the capacitance is again just the oxide capacitance,

$$C'(inv) = C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \qquad 3.33$$

The ideal C-V characteristics of the MOS capacitor with a p-type substrate are shown in Figure 3.22.The three dashed segments correspond to the three components C_{ox}, C_{SD}' and C_{min}'. The solid curve is the ideal net capacitance of the MOS capacitor. Moderate inversion, which is indicated in the figure, is the transition region between the point when only the space charge density changes with gate voltage and when only the inversion charge density changes with gate voltage.

The point on the curve that corresponds to the flat-band condition, occurs between the

accumulation and depletion conditions. The capacitance at flat-band is given by,

$$C_{FB}' = \frac{\varepsilon_{ox}}{t_{ox} + \left(\frac{\varepsilon_{ox}}{\varepsilon_s}\right)\sqrt{\left(\frac{kT}{e}\right)\left(\frac{\varepsilon_s}{eN_a}\right)}} \qquad 3.34$$

The flat-band capacitance is a function of oxide thickness as well as semiconductor doping. The general location of this point on the C-V plot is indicated in Figure 3.22.

The same type of ideal C-V characteristics are obtained for a MOS capacitance with an n-type substrate by changing the sign of the voltage axis. The accumulation condition is obtained for a positive gate bias and the inversion condition is obtained for a negative gate bias. This ideal curve is shown in Figure 3.23.

3.2.2 Frequency Effects

The MOS capacitor with a p-type substrate and biased in the inversion condition was shown in Figure 3.21a. The previous argument was that a differential change in the capacitor voltage in the ideal case causes a differential change in the inversion layer charge density. The source of the electrons that produces a change in the inversion charge density will now be considered.

There are two sources of electrons that can change the charge density of the inversion layer. The first source is by diffusion of minority carrier electrons from the p-type substrate across the space charge region. This diffusion process is the same as that in a reverse-biased pn junction which generates the ideal reverse saturation current. The second source of electrons is by thermal generation of electron-hole pairs within the space charge region. This process is again the same as that in a reverse-biased pn junction which generates the reverse-biased generation current. Both of these processes generate electrons at a particular rate. The electron concentration in the inversion layer, then cannot change instantaneously. If the ac voltage across the MOS capacitor changes rapidly, the change in the inversion layer charge will not be able to respond. The C-V characteristics will then be a function of the frequency of the ac signal used to measure the capacitance.

In the limit of a very high frequency, the inversion layer charge will not respond to a differential change in capacitor voltage. Figure 3.24 shows the charge distribution in the MOS capacitor with a p-type substrate. At a high signal frequency, the differential change occurs at the metal and in the space charge width in the semiconductor. The capacitance of the MOS capacitor is then C_{min}'. The high-frequency and low-frequency limits of the C-V characteristics are shown in Figure 3.25. In general, high frequency corresponds to a value on the order of 1MHz and low frequency corresponds to values in the range of 5 to 100 Hz

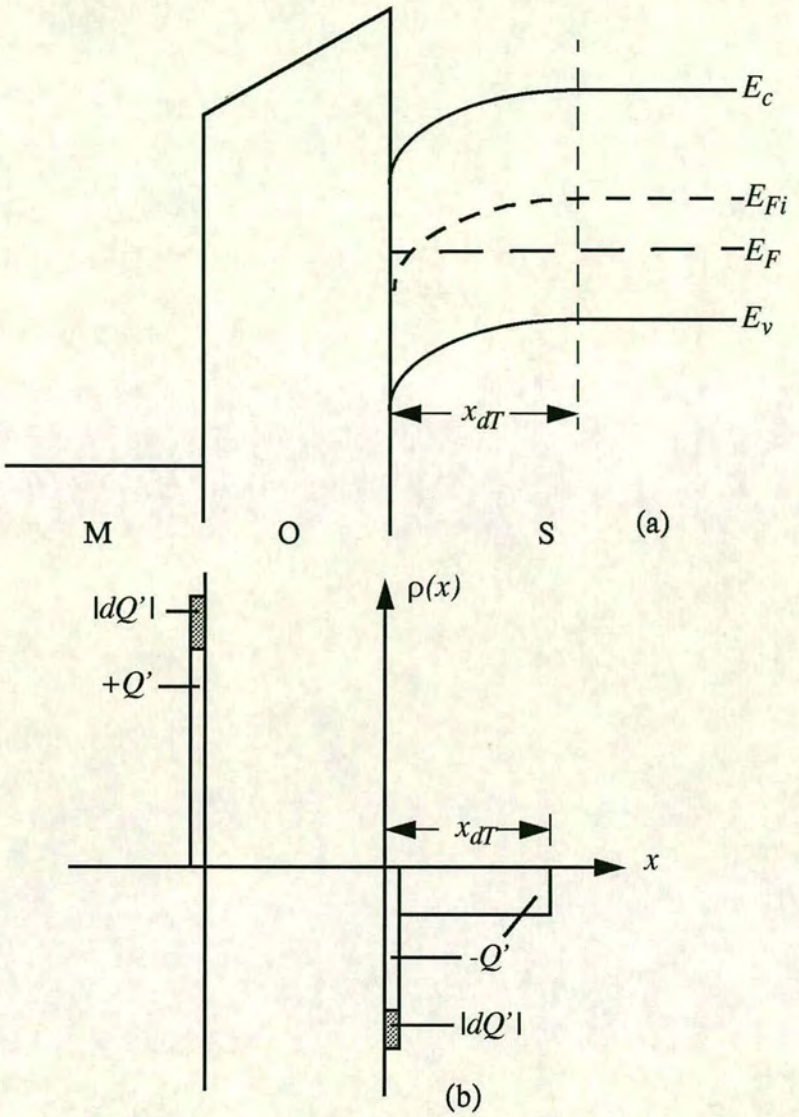


Figure 3.21(a) Energy-band diagram through a MOS capacitor at inversion. (b) Differential charge distribution at inversion for a low-frequency differential change in gate voltage.

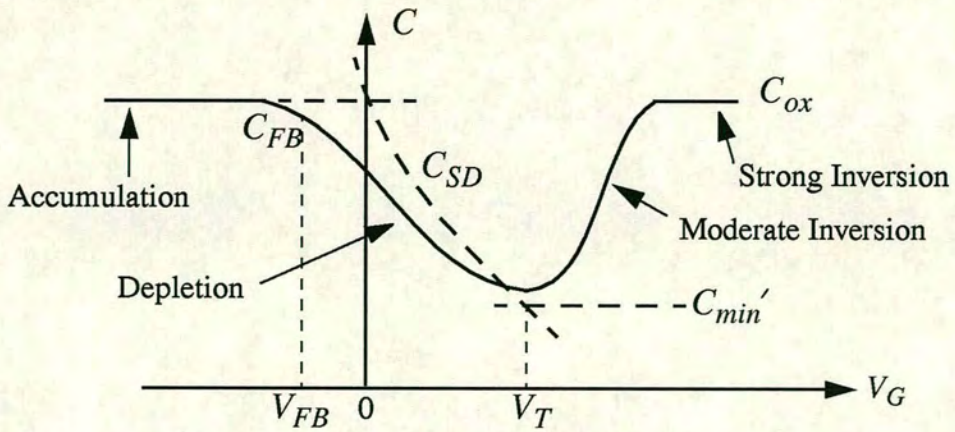


Figure 3.22 Ideal low-frequency capacitance verses gate voltage of a MOS capacitor with a p-type substrate. Individual capacitance components are also shown

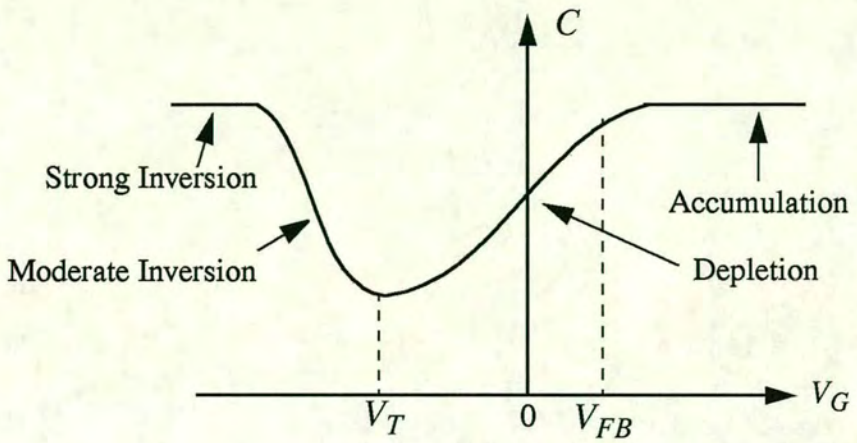


Figure 3.23 Ideal low-frequency capacitance verses gate voltage of a MOS capacitor with an n-type substrate.

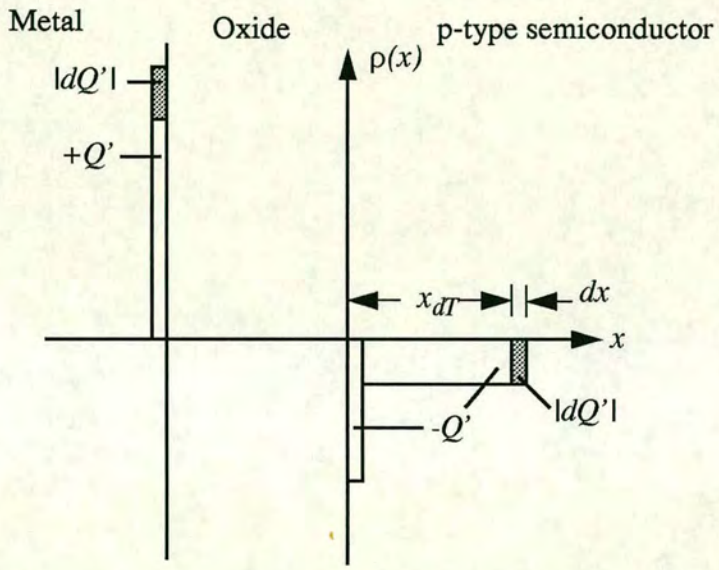


Figure 3.24 Differential charge distribution at inversion for a high-frequency differential change in gate voltage.

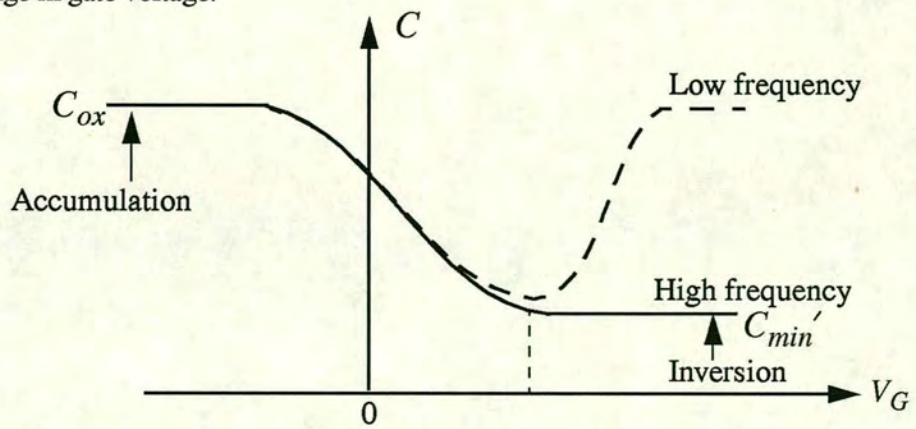


Figure 3.24 Low-frequency and high-frequency capacitance verses gate voltage of a MOS capacitor with a p-type substrate.

3.2.3 Fixed Oxide and Interface Charge Effects

The fixed oxide charge was previously shown to affect the threshold voltage. This charge also affects the flat-band voltage. The flat-band voltage was previously given as,

$$V_{FB} = \phi_{ms} - \frac{Q_{ss}'}{C_{ox}} \qquad 3.18$$

where Q_{ss}' is the equivalent fixed oxide charge and $\phi_{ms}$is the metal-semiconductor work function difference. The flat-band voltage shifts to more negative voltages for a positive fixed oxide charge. Since the oxide charge is not a function of gate voltage, the curves show a parallel shift with oxide charge and the shape of the C-V curves remains as the ideal characteristics. Figure 3.26 shows the high-frequency characteristics of a MOS capacitor with a p-type substrate for several values of fixed positive oxide charge.

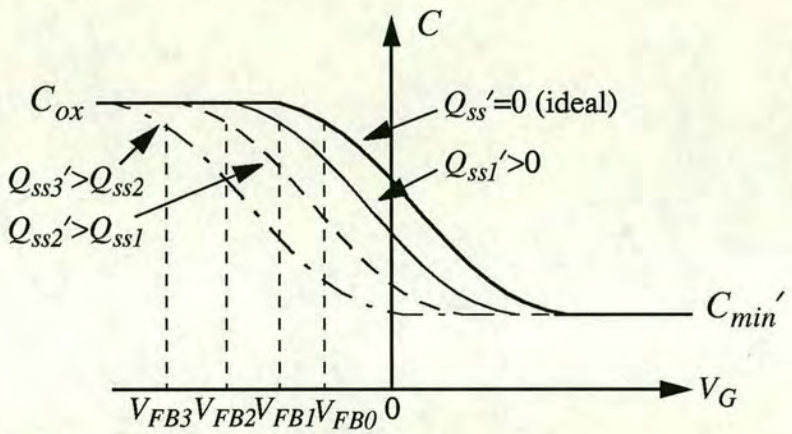


Figure 3.26 High-frequency capacitance versus gate voltage of a MOS capacitor with a p-type substrate for several values of effective trapped oxide charge.

The C-V characteristics can be used to determine the equivalent fixed oxide charge. For a given MOS structure, ϕ_{ms} and C_{ox} are known so the ideal flat-band voltage and flat-band capacitance can be calculated. The experimental value of flat-band voltage can be measured from the C-V curve and the value of fixed oxide charge can then be determined. The C-V measurements are a valuable diagnostic tool to characterize a MOS device.

Figure 3.27 shows the energy-band diagram of a semiconductor at the oxide-semiconductor interface. The periodic nature of the semiconductor is abruptly terminated at the interface so that allowed electronic energy levels will exist within the forbidden bandgap. These allowed energy states are referred to as interface states. Charge can flow between the semiconductor

and interface states, in contrast to the fixed oxide charge. The net charge in these interface states is a function of the position of the Fermi level in the bandgap.

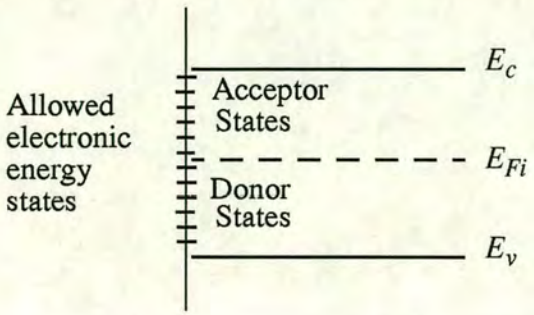


Figure 3.27 Schematic diagram showing interface states at the oxide-semiconductor interface.

In general, acceptor states exist in the upper half of the bandgap and donor states exist in the lower half of the bandgap. An acceptor state is neutral if the Fermi level is below the state, and becomes negatively charged if the Fermi level is above the state. A donor state is neutral is the Fermi level is above the state and becomes positively charged if the Fermi level is below the state. The charge of the interface states is then a function of the gate voltage applied across the MOS capacitor.

Figure 3.28a shows the energy-band diagram in a p-type semiconductor of a MOS capacitor biased in the accumulation condition. In this case, there is a net positive charge trapped in the donor states. If now, the gate voltage is changed to produce the energy-band diagram in Figure 3.28b. The Fermi level corresponds to the intrinsic Fermi level at the surface and the interface states are all neutral. This particular bias condition is known as *midgap*. Figure 3.28c shows the condition at inversion in which there is now a net negative charge in the acceptor states.

The net charge in the interface states changes from positive to negative as the gate voltage sweeps from the accumulation, depletion, to the inversion condition. The C-V curves previously shifted in the negative gate voltage direction due to fixed positive oxide charges. When interface states are present, the amount and direction of the shift changes as the gate voltage sweeps because the amount and sign of the interface trapped charge changes. This has the affect of smearing out the C-V plot as seen in Figure 3.29. The amount of smearing out can be used to determine the density of interface states

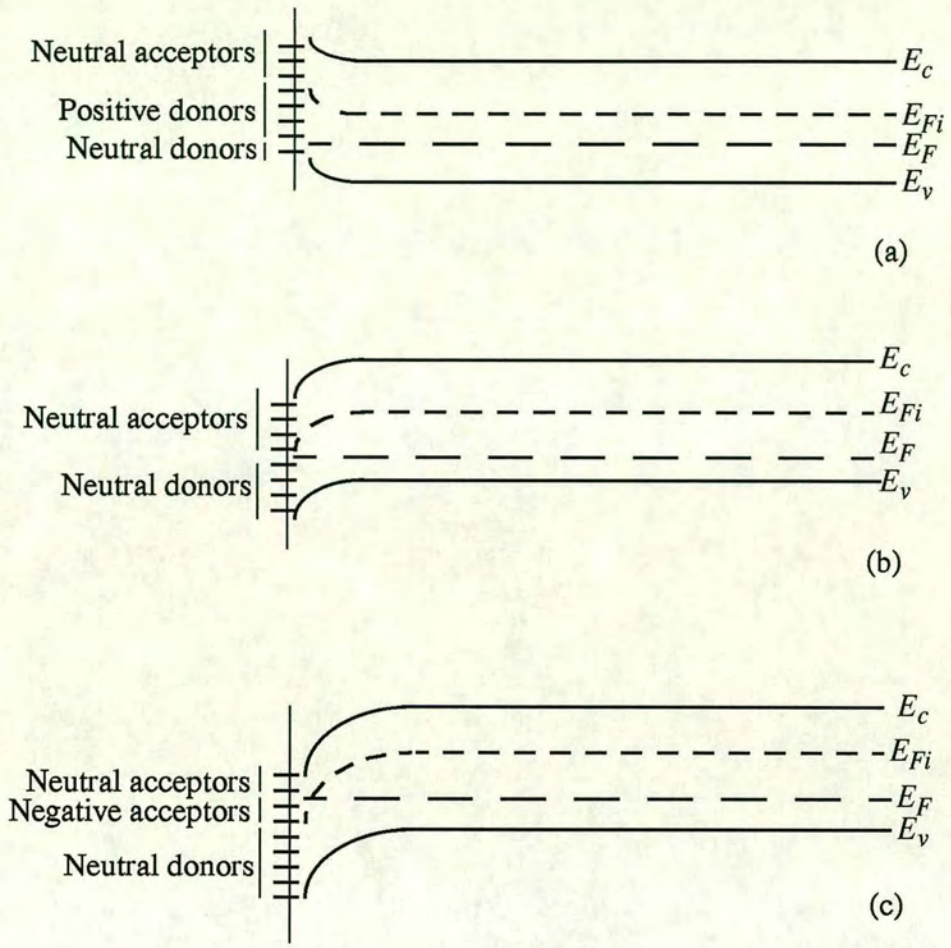


Figure 3.28 Energy-band diagram in a p-type semiconductor showing the charge trapped in the interface states when the MOS capacitor is biased (a) in accumulation, (b) at midgap, and (c) at inversion.

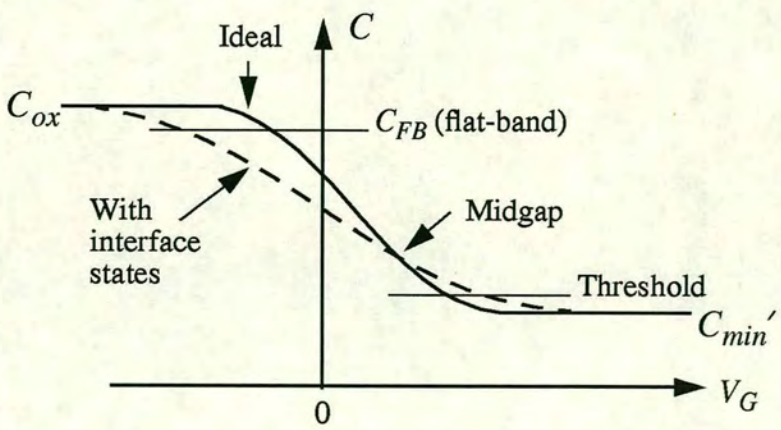


Figure 3.29 High-frequency C-V characteristics of a MOS capacitor showing effects of interface states.

3.3 The Basic MOSFET Operation [31]

3.3.1 MOSFET Structures

The current in a MOS field-effect transistor is due to the flow of charge in the inversion layer or channel region adjacent to the oxide-semiconductor interface [32].

There are four basic MOSFET device types. Figure 3.30 shows an n-channel enhancement mode MOSFET. In the enhancement mode, the semiconductor substrate is not inverted directly under the oxide with zero gate voltage. A positive voltage induces the electron inversion layer, which then 'connects' the n-type source and the n-type drain regions. The source terminal is the source of carriers that flow through the channel to the drain terminal. For this n-channel device, electrons flow from the source to the drain so the conventional current will enter the drain and leave the source. The conventional circuit symbol for this n-channel enhancement mode device is also shown in this figure. Figure 3.31 shows an n-channel depletion mode MOSFET. An n-channel region exists under the oxide with zero volts applied to the gate, this is due to the p-type substrate having a negative threshold voltage. The n-channel shown in this figure can be an electron inversion layer or an intentionally doped n-region. The conventional circuit symbol for this device is also shown in the figure. Figures 3.32a and 3.32b show a p-channel enhancement mode MOSFET and a p-channel depletion mode MOSFET respectively. In the p-channel enhancement mode device, a negative gate voltage must be applied to create an inversion layer of holes that will 'connect' the p-type source and drain regions. Holes flow from the source to the drain so the conventional current will enter the source and leave the drain. A p-channel region exists in the depletion mode device even with zero gate voltage. The conventional circuit symbols are shown in the figures.

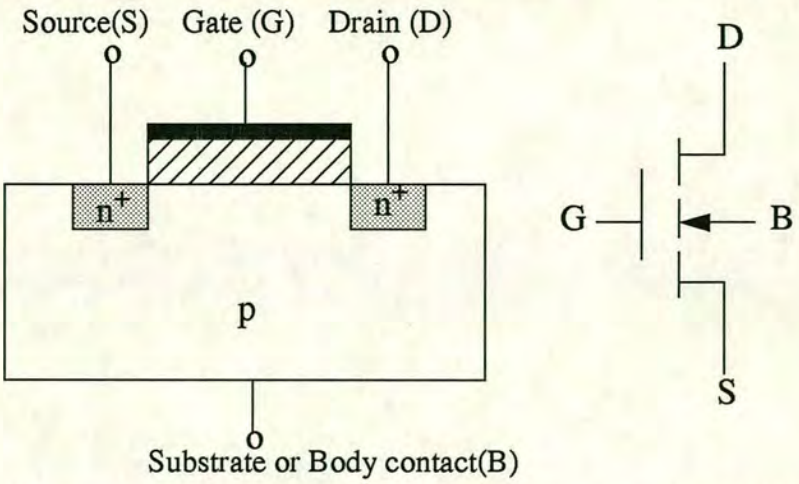


Figure 3.30 Cross section and circuit symbol for an enhancement mode n-MOSFET.

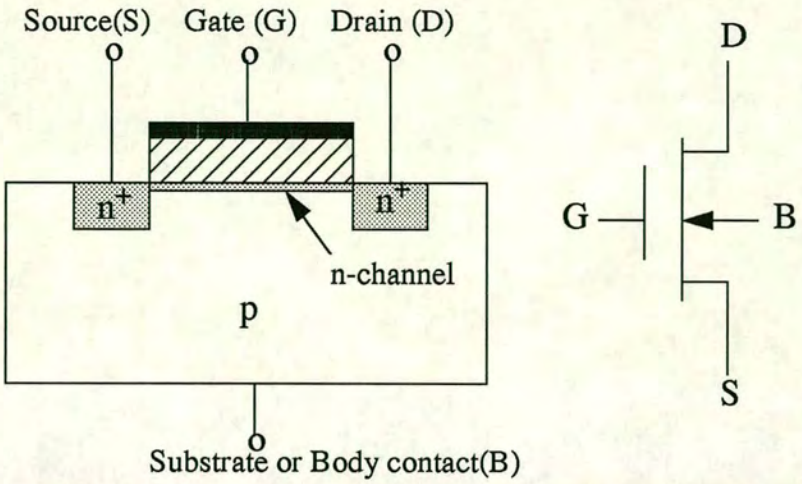


Figure 3.31 Cross section and circuit symbol for a depletion mode n- MOSFET.

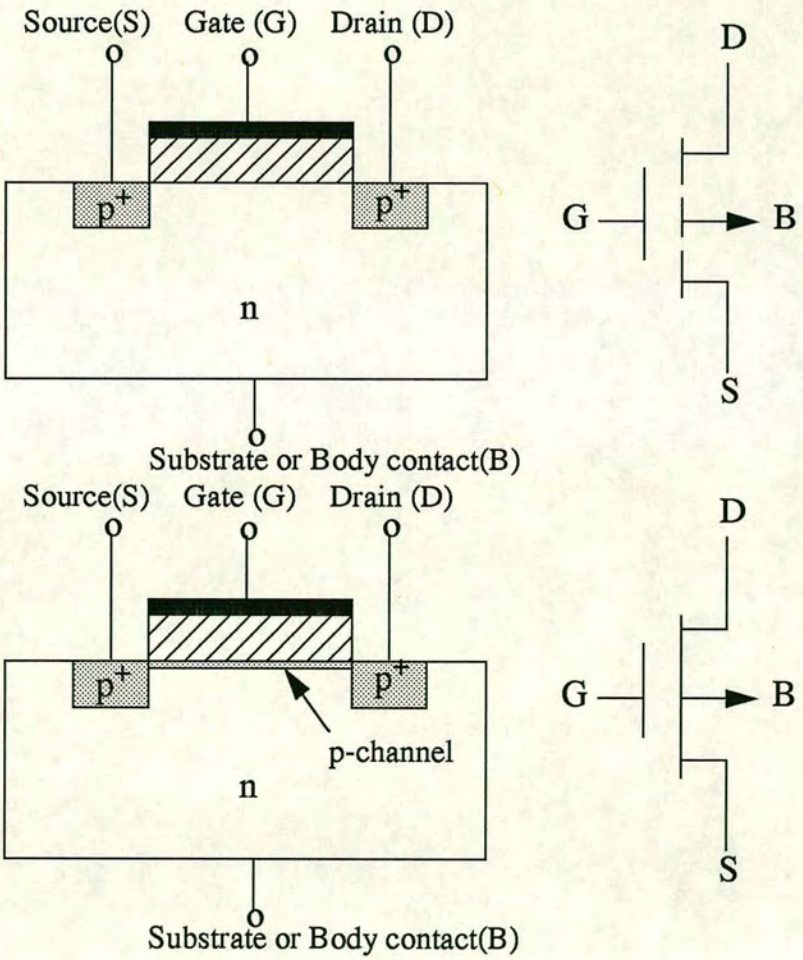


Figure 3.32 Cross section and circuit symbol for (a) a p-channel enhancement mode MOSFET, and (b) a p-channel depletion mode MOSFET.

3.3.2 Current-Voltage Characteristics

Figure 3.33a shows an n-channel enhancement mode MOSFET with a gate to source voltage, V_{GS}, that is less than the threshold voltage and with only a very small drain to source voltage, V_{DS}. The source and substrate, or body, terminals are held at ground potential. With this bias configuration, there is no electron inversion layer, the drain to substrate pn junction is reverse biased, and apart from the pn junction leakage currents, the drain current is zero.

Figure 3.33b shows the same MOSFET with a applied gate voltage such that $V_{DS} > V_T$. An electron inversion layer has now been created and when a small drain voltage is applied, the electrons in the inversion layer will flow from the source to the positive drain terminal. The conventional current enters the drain terminal and leaves the source terminal. In this ideal case, there is no current through the oxide to the gate terminal.

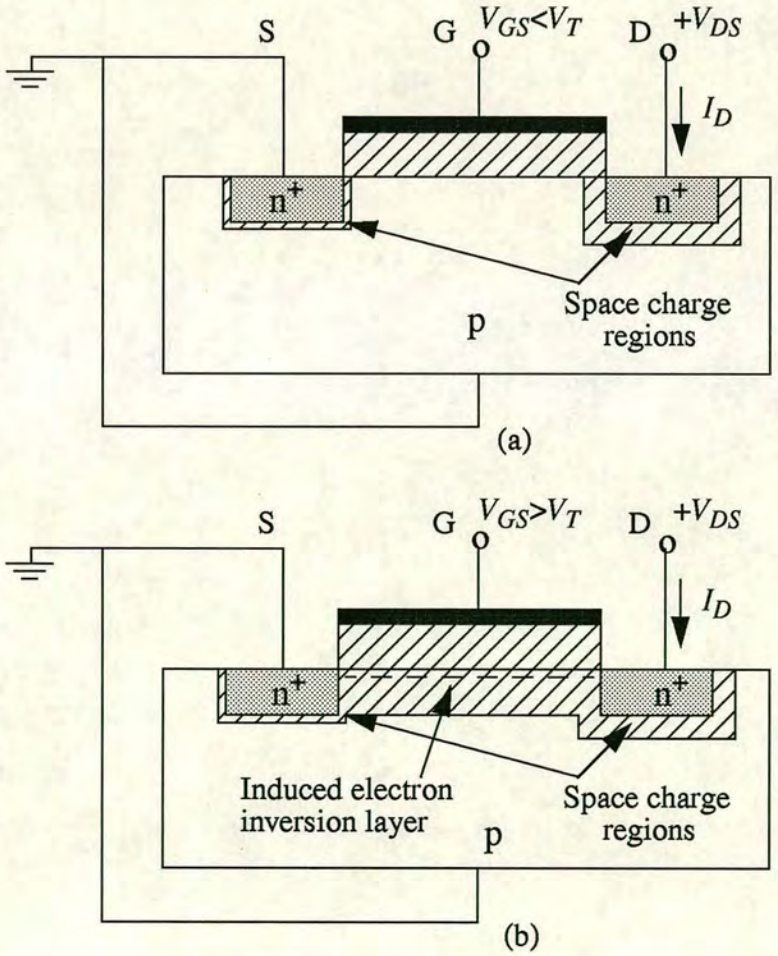


Figure 3.33 The n-channel enhancement mode MOSFET (a) with an applied gate voltage $V_{GS}<V_T$, and (b) with an applied gate voltage $V_{GS}>V_T$.

For small V_{DS} values, the channel region has the characteristics of a resistor,

$$I_D = g_d V_{DS} \qquad 3.35$$

where g_d is defined as the channel conductance in the limit as $V_{DS} \rightarrow 0$. The channel conductance is given by

$$g_d = \frac{W}{L}\mu_n |Q_n'| \qquad 3.36$$

where μ_n is the mobility of the electrons in the inversion layer and $|Q_n'|$ is the magnitude of the inversion layer charge per unit area. The inversion layer charge is a function of the gate voltage; thus the basic MOS transistor action is the modulation of the channel conductance by the gate voltage. The channel conductance, in turn, determines the drain current. For the time being, the mobility is assumed to be a constant.

The I_D verses V_{DS} characteristics, for small values of V_{DS}, are shown in Figure 3.34

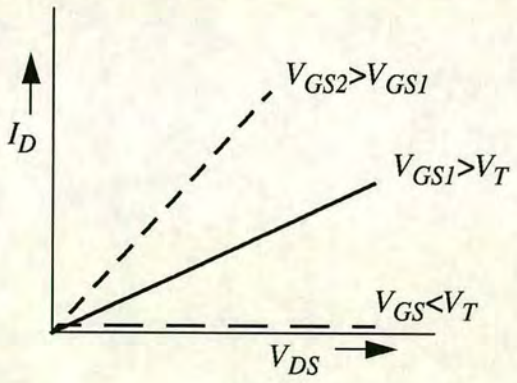


Figure 3.34 I_D verses V_{DS} characteristics for small values of V_{DS} at three V_{GS} voltages.

When $V_{GS} < V_T$, the drain current is zero. As V_{GS} becomes larger than V_T, the channel inversion charge density increases, which increases the channel conductance. A larger value of g_d produces a larger initial slope of the I_D verses V_{DS} characteristic as shown in the figure.

Figure 3.35a shows the basic MOS structure for the case when $V_{GS} > V_T$ and the applied V_{DS} is small. The thickness of the inversion channel layer in the figure qualitatively indicates the relative charge density, which is essentially constant along the entire channel length for this case. The corresponding I_D verses V_{DS} curve is shown in the figure.

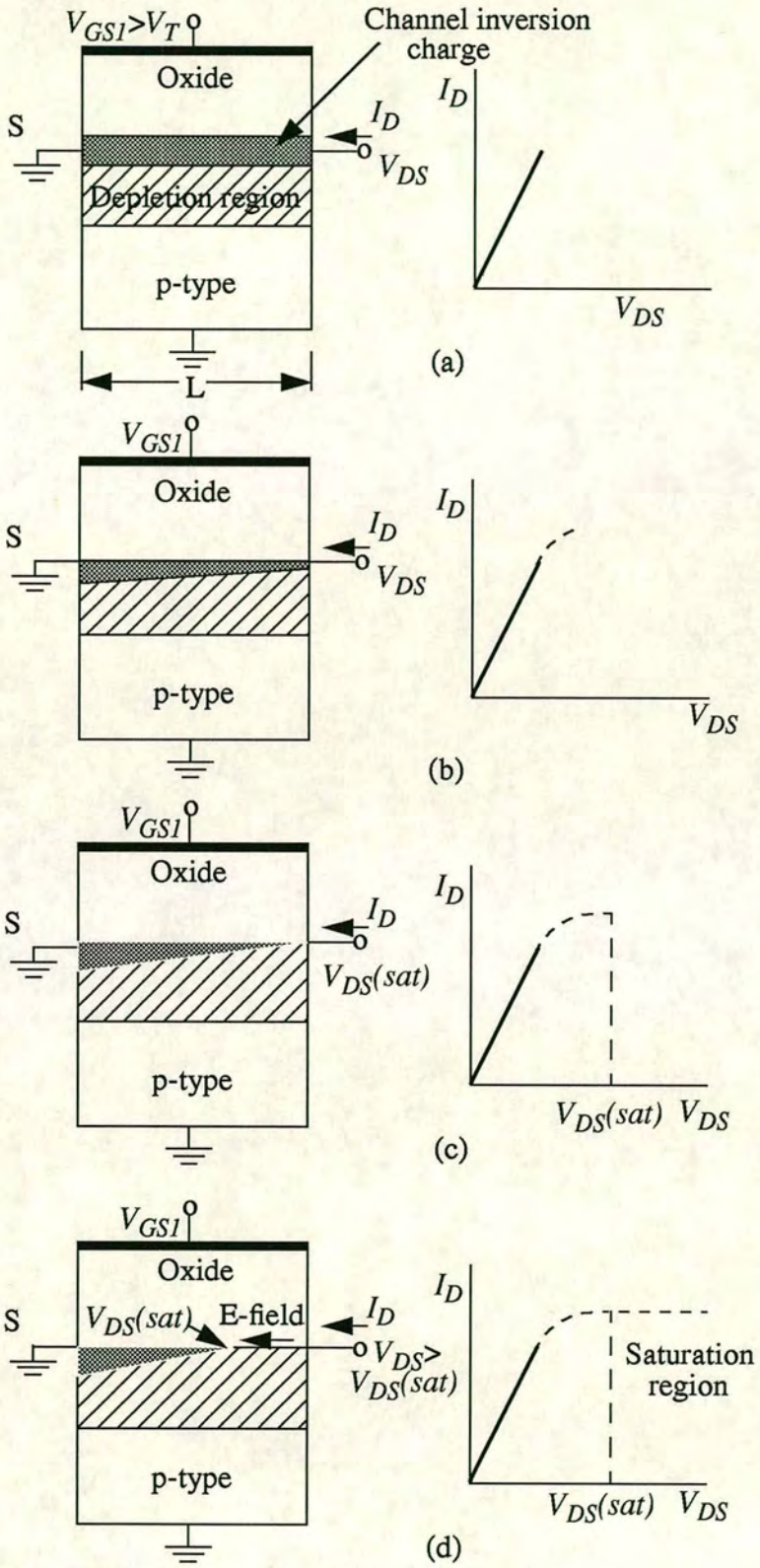


Figure 3.35 Cross section and I_D verses V_{DS} curve when $V_{GS}>V_T$ for (a) a small V_{DS} value, (b) a larger V_{DS} value, (c) a value of $V_{DS}=V_{DS}$(sat), and (d) a value of $V_{DS}>V_{DS}$(sat).

Figure 3.35b shows the situation when the V_{DS} value increases. As the drain voltage increases, the voltage drop across the oxide near the drain terminal decreases. The incremental conductance of the channel at the drain decreases which then means that the slope of the I_D verses V_{DS} curve will decrease.

When V_{DS} increases to the point where the potential drop across the oxide at the drain terminal is equal to V_T, the induced inversion charge density is zero at the drain terminal. This effect is schematically shown in Figure 3.35c. At this point, the incremental conductance at the drain is zero which means that the slope of the I_D verses V_{DS} curve is zero. Then,

$$V_{GS} - V_{DS}(sat) = V_T \qquad 3.37a$$

or

$$V_{DS}(sat) = V_{GS} - V_T \qquad 3.37b$$

where $V_{DS}(sat)$ is the drain to source voltage producing zero inversion charge density at the drain terminal.

When V_{DS} becomes larger than the $V_{DS}(sat)$ value, the point in the channel at which the inversion charge is zero moves towards the source terminal. In this case, electrons enter the channel at the source, travel through the channel towards the drain, and then at the point where the charge goes to zero, the electrons are injected into the space charge region where they are swept by the *E*-field to the drain contact. If the change in channel length ΔL is small compared to the original length L (i.e. for long-channel MOSFET devices), the drain current will be a constant for $V_{DS} > V_{DS}(sat)$. This region of operation is shown in Figure 3.35d.

When V_{GS}changes, the I_D verses V_{DS} curve will change, the initial slope of I_D verses V_{DS} increases. From Equation 3.37b, the value of $V_{DS}(sat)$ is a function of V_{GS}. The family of curves for this n-channel enhancement mode MOSFET is shown in Figure 3.36

Figure 3.37 shows an n-channel depletion mode MOSFET. If the n-channel region is actually an induced electron inversion layer created by the metal-semiconductor work function difference and the fixed charge in the oxide, the current-voltage characteristics are the same as that for the enhancement mode device except that V_T is a negative quantity.

For the case when the n-channel region is actually an n-type semiconductor, a negative gate voltage will induce a space charge region under the oxide, reducing the thickness of the n-

channel region. The reduced thickness decreases the channel conductance, which reduces the drain current. A positive gate voltage will create an electron accumulation layer, which increases the drain current. The channel thickness t_c must be less than the maximum induced space charge width in order to be able to turn the device off. The general I_D verses V_{DS} family of curves for an n-channel depletion mode MOSFET are given in Figure 3.38.

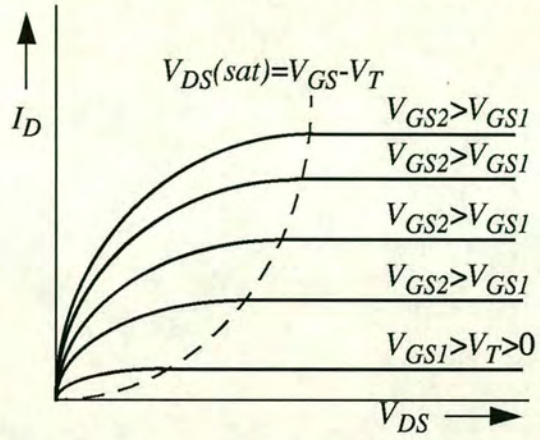


Figure 3.36 Family of I_D verses V_{DS}curves for an n-channel enhancement mode MOSFET.

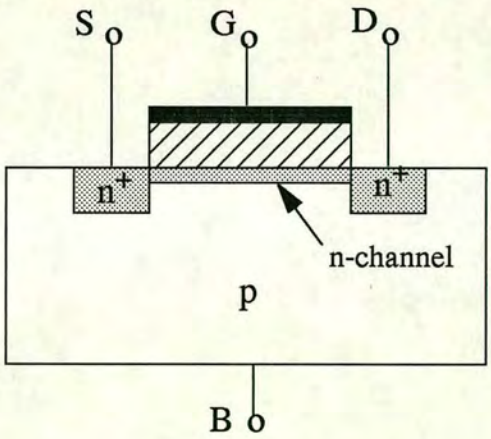


Figure 3.37 Cross section of an n-channel mode MOSFET.

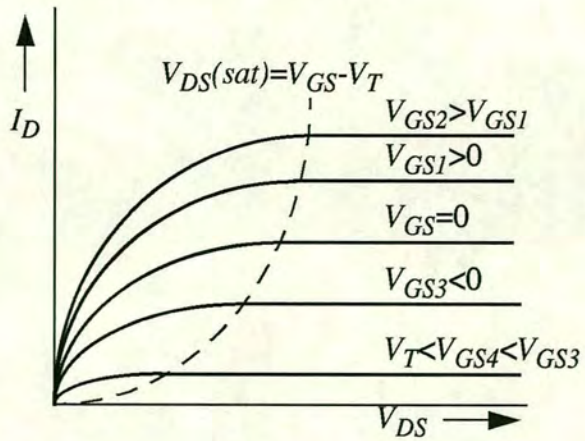


Figure 3.38 Family of I_D verses V_{DS} curves for an n-channel depletion mode MOSFET.

3.3.3 Current-Voltage Mathematical Derivations

In the previous section, the current-voltage characteristics were qualitatively discussed. In this section, the mathematical relationships between drain current, gate to source voltage, and drain to source voltage will be derived.

Figure 3.39 shows the geometry of the device that will be used in the following derivations. The following assumptions are made:

1. The current in the channel is due to drift rather than diffusion.
2. There is no current through the oxide.
3. A gradual channel approximation is used in which the lateral electric field E_x is constant or $\frac{\partial E_y}{\partial y} \gg \frac{\partial E_x}{\partial x}$
4. Any fixed oxide charge is assumed to be an equivalent charge density at the oxide-semiconductor interface.
5. The carrier mobility in the channel is constant.

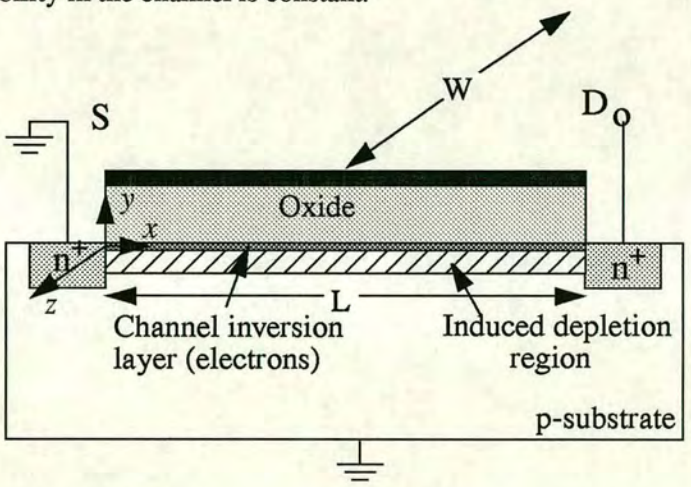


Figure 3.39 Geometry of a MOSFET for I_D verses V_{DS} derivation.

Ohm's Law can be written as

$$J_x = \sigma E_x \qquad 3.38$$

where σ is the channel conductivity and E_x is the electric field along the channel created by the drain to source voltage. The channel conductivity is given below where μ_n is the electron mobility and $n(y)$ is the electron concentration in the inversion layer.

$$\sigma = e\mu_n n(y) \qquad 3.39$$

The total channel current is found by integrating J_x over the cross-sectional area in the y and z directions. Then

$$I_x = \iint_{y\,z} J_x dy dz \qquad 3.40$$

given that

$$Q_n' = -\int en(y)dy \qquad 3.41$$

where Q_n' is the inversion layer charge per unit area and is a negative quantity for this case.

Equation 3.40 then becomes

$$I_s = -W\mu_n Q_n' E_x \qquad 3.42$$

where W is the channel width, the integration over z.

The two concepts of charge neutrality and Gauss's law will be used in the current-voltage derivation. Figure 3.40 shows the charge densities through the device for $V_{GS} > V_T$. The charges are all given in terms of charge per unit area.

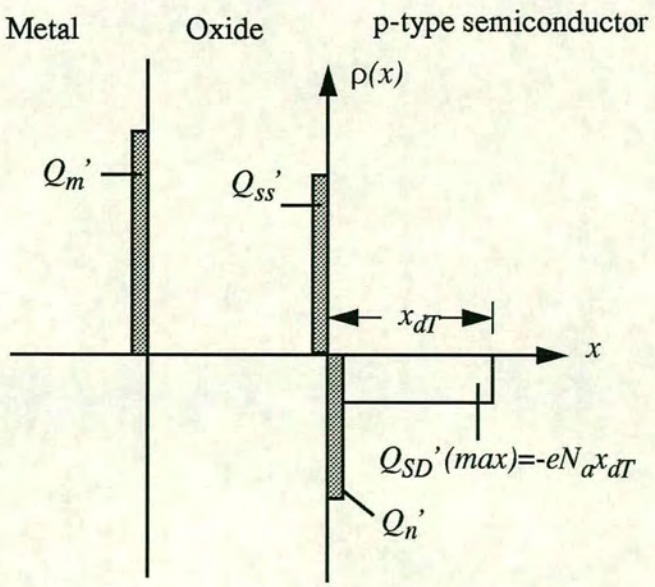


Figure 3.40 Charge distribution in the n-channel enhancement mode MOSFET for $V_{GS} > V_T$.

Using the concept of charge neutrality,

$$Q_m' + Q_{ss}' + Q_n' + Q_{SD}'(max) = 0 \qquad 3.43$$

The inversion layer charge and induced space charge are negative for this device. Gauss's law can be written as

$$\oint_S \varepsilon E_n dS = Q_T \qquad 3.44$$

where the integral is over a closed surface, Q_T is the total charge enclosed by the surface, and E_n is the outward directed normal component of the electric field crossing the surface S. Gauss's law will be applied to the surface defined in Figure 3.41.

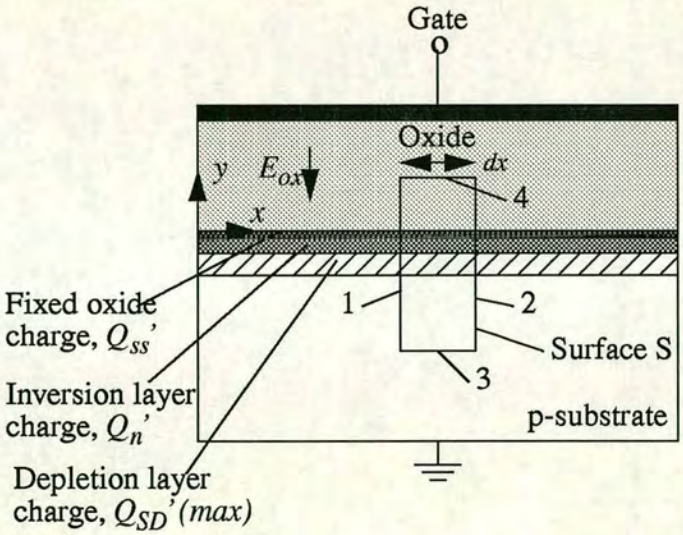


Figure 3.41 Geometry for applying Gauss's law.

Since the surface must be enclosed, the two end surfaces in the x-y plane must be taken into account. These two surfaces however, have no z component and so do not contribute to the integral of Equation 3.44.

Considering the surfaces labelled 1 and 2 in Figure 3.42, the assumption that the lateral electric field is constant means that E_x into surface 2 is the same as E_x out of surface 1. Since the integral in Equation 3.44 involves the outward component of the E-field, the contributions of surfaces 1 and 2 cancel each other. Surface 3 is in the neutral p-region so the electric field is zero at this surface.

Surface 4 is the only surface that contributes to Equation 3.44. Taking into account the direction of the electric field in the oxide, Equation 3.44 becomes

$$\oint_S \varepsilon E_n dS = -\varepsilon_{ox} E_{ox} W dx = Q_T \qquad 3.45$$

where ε_{ox} is the permittivity of the oxide. The total charge enclosed is

$$Q_T = (Q_{ss}' + Q_n' + Q_{SD}'(max))Wdx \qquad 3.46$$

Combining Equations 3.45 and 3.46,

$$-\varepsilon_{ox} E_{ox} = Q_{ss}' + Q_n' + Q_{SD}'(max) \qquad 3.47$$

Figure 3.43a shows the oxide and channel, the source is at ground potential. The voltage V_x is the potential in the channel at a point x along the channel length. The potential difference across the oxide at x is a function of V_{GS}, V_x, and the metal-semiconductor work function difference.

The energy-band diagram through the MOS structure at point x is shown in Figure 3.43b. The Fermi level in the p-type semiconductor is E_{Fp} and the Fermi level in the metal is E_{Fm}. So that

$$E_{Fp} - E_{Fm} = e(V_{GS} - V_x) \qquad 3.48$$

Considering the potential barriers gives,

$$V_{GS} - V_x = (\phi_m' + V_{ox}) - \left(\chi' + \frac{E_g}{2e} - \phi_s + \phi_{fp}\right) \qquad 3.49$$

which can also be written as,

$$V_{GS} - V_x = V_{ox} + 2\phi_{fp} + \phi_{ms} \qquad 3.50$$

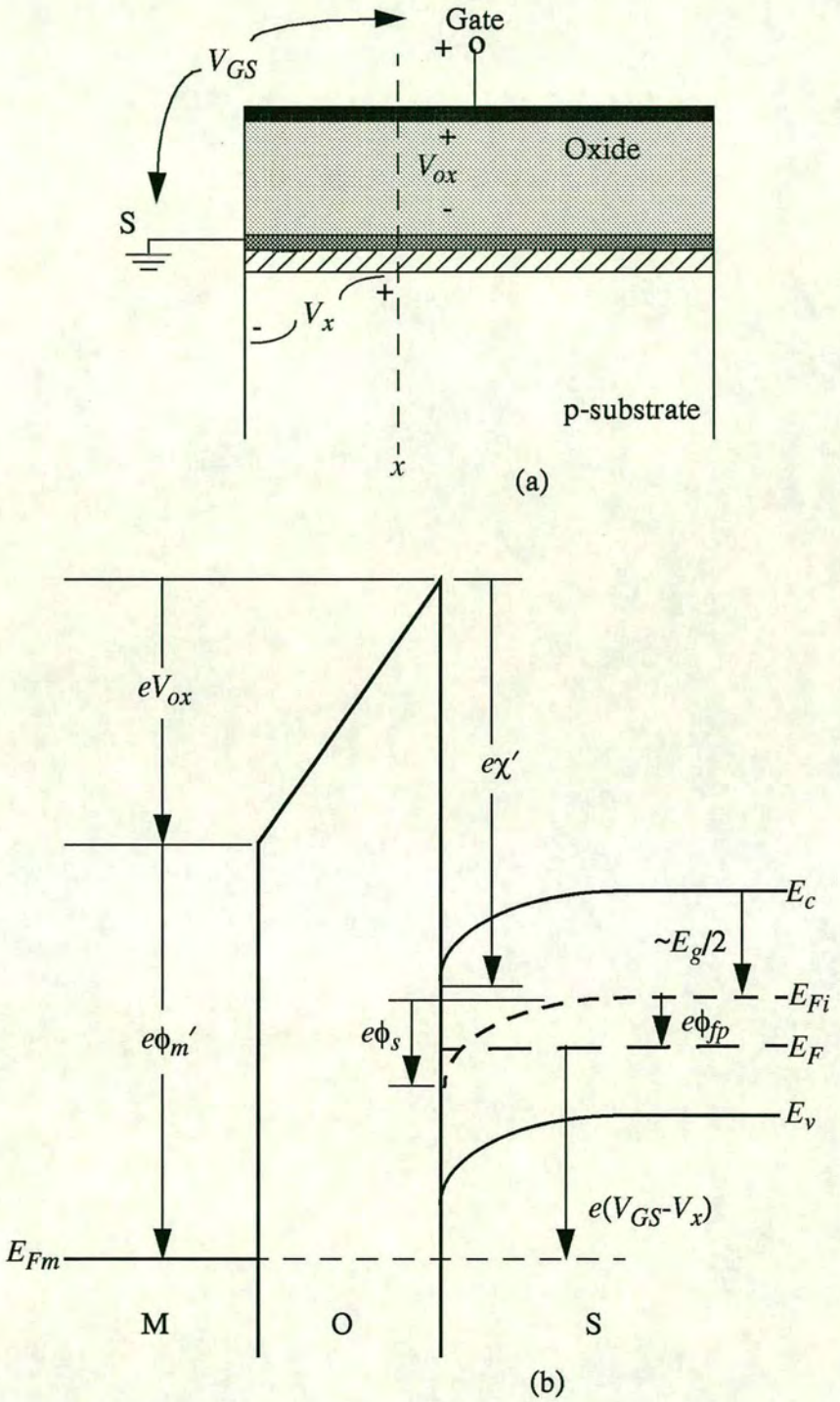


Figure 3.43 (a) Potentials at a point x along the channel. (b) Energy-band diagram through the MOS structure at the point x.

where ϕ_m is the metal-semiconductor work function difference, and $\phi_s = 2\phi_{fp}$ for the inversion condition.

The electric field in the oxide is

$$E_{ox} = \frac{V_{ox}}{t_{ox}} \qquad 3.51$$

Combining Equations 3.47, 3.50 and 3.51,

The inversion charge density, Q_n', from Equation 3.52 can be substituted into Equation 3.42 to give

$$I_x = -W\mu_n C_{ox}\frac{dV_x}{dx}\{(V_{GS} - V_x) - V_T\} \qquad 3.53$$

where ε_x=-dV_x/dx and V_T is the threshold voltage defined by Equation 3.24.

Integrating Equation 3.53 over the length of the channel gives

$$\int_0^L I_x dx = -W\mu_n C_{ox}\int_{V_x(0)}^{V_x(L)}[(V_{GS} - V_T) - V_x]dV_x \qquad 3.54$$

The mobility μ_n is assumed to be constant. For the n-channel device, the drain current enters the drain terminal and is a constant along the entire channel length. Letting I_D=-I_x, Equation 3.54 becomes

$$I_D = \frac{W\mu_n C_{ox}}{2L}[2(V_{GS} - V_T)V_{DS} - V_{DS}^2] \qquad 3.55$$

Equation 3.55 is valid for $V_{GS} \geq V_T$ and for $0 \leq V_{DS} \leq V_{DS}(sat)$.

Figure 3.44 shows plots of Equation 3.55 as a function of $V_{DS}$for several values of V_{GS}. We can find the value of V_{DS}at the peak current value from $\partial I_D/\partial V_{DS} = 0$. Then using Equation 3.55, the peak current occurs when

$$V_{DS} = V_{GS} - V_T \qquad 3.56$$

This value of V_{DS} is just $V_{DS}(sat)$, the point at which saturation occurs. For $V_{DS}>V_{DS}(sat)$, the ideal drain current is a constant and is equal to

$$I_D(sat) = \frac{W\mu_n C_{ox}}{2L}[2(V_{GS} - V_T)V_{DS}(sat) - V_{DS}^2(sat)] \qquad 3.57$$

Using Equation 3.56 for $V_{DS}(sat)$, Equation 3.57 becomes

$$I_D(sat) = \frac{W\mu_n C_{ox}}{2L}(V_{GS} - V_T)^2 \qquad 3.58$$

Equation 3.55 is the ideal current-voltage relationship of the n-channel MOSFET in the non-saturation region for $0 \le V_{DS} \le V_{DS}(sat)$ and Equation 3.58 is the ideal current-voltage relationships the n-channel MOSFET in the saturation region for $V_{DS}>V_{DS}(sat)$. These I-V expressions are explicitly derived for n-channel enhancement mode devices. The same equations also apply to the n-channel depletion mode MOSFET where the threshold voltage V_T is a negative quantity.

The mobility and threshold voltage can be experimentally determined using the I-V relations. From Equation 3.55, for very small values of V_{DS},

$$I_D = \frac{W\mu_n C_{ox}}{L}(V_{GS} - V_T)V_{DS} \qquad 3.59a$$

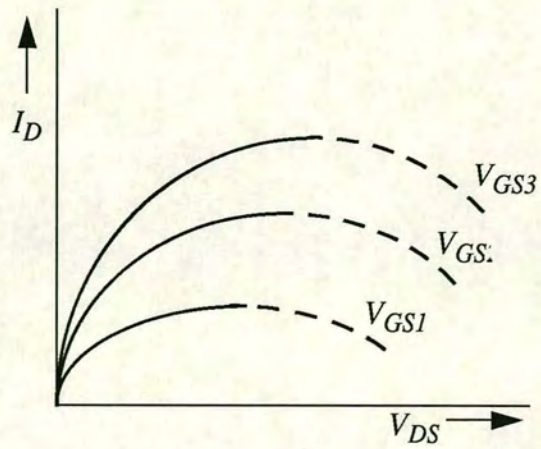


Figure 3.44 Plots of I_D verses V_{DS} from Equation 3.55.

Figure 3.45a shows a plot of Equation 3.59a as a function of V_{GS} for constant V_{DS}. A straight line is fitted through the points. The deviation from the straight line at low values of V_{GS} is due to subthreshold conduction and the deviation at higher values of V_{GS} is due to mobility being a function of gate voltage. More will be said about these effects later. The extrapolation of the straight line to zero current gives the threshold voltage and the slope is proportional to the inversion carrier mobility.

If we take the square root of Equation 3.58, we obtain,

$$\sqrt{I_D(sat)} = \sqrt{\frac{W\mu_n C_{ox}}{2L}}(V_{GS} - V_T) \qquad 3.59b$$

Figure 3.45b is a plot of Equation 3.59b. In the ideal case, the same information is obtained for both curves, however for short channel devices, the threshold voltage may be a function of V_{DS}. Since Equation 3.59b applies to devices biased in the saturation region, the V_T parameter in this equation may differ from the extrapolated value determined in Figure 3.45a. In general, the non-saturation current-voltage characteristics will produce the more reliable extrapolation. The voltage polarities and current direction are the reverse of those in the n-channel device. For the current direction shown in the figure, the I-V relations for the p-channel MOSFET are,

$$I_D = \frac{W\mu_p C_{ox}}{2L}[2(V_{SG} + V_T)V_{SD} - V_{SD}^2] \qquad 3.60$$

for $0 \le V_{SD} \le V_{SD}(sat)$, and

$$I_D(sat) = \frac{W\mu_p C_{ox}}{2L}(V_{SG} + V_T)^2 \qquad 3.61$$

for $V_{DS} > V_{DS}(sat)$, where

$$V_{SD}(sat) = V_{SG} + V_T \qquad 3.62$$

The current-voltage relationship of a p-channel device can be obtained using the same type of analysis. Figure 3.46 shows a p-channel enhancement mode MOSFET.

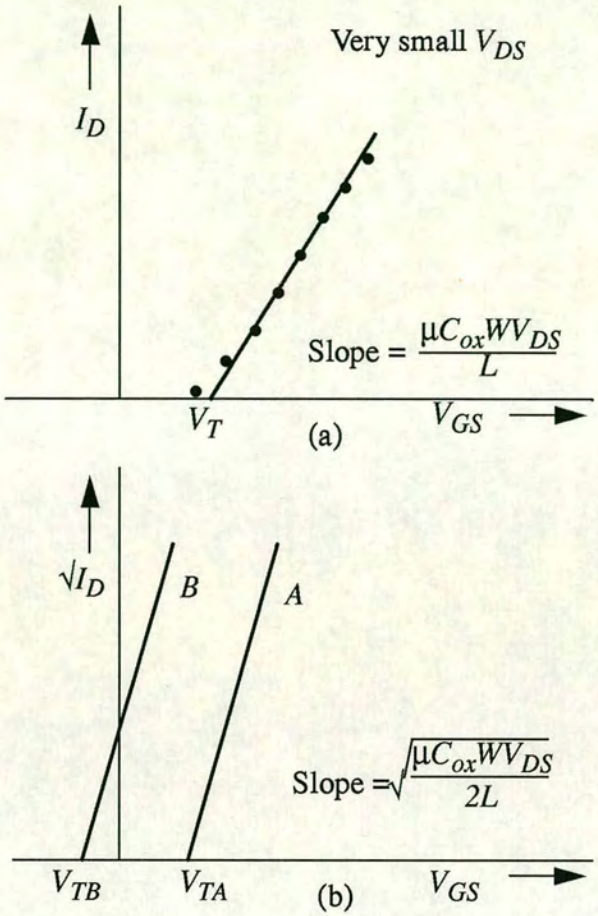


Figure 3.45 (a) I_D verses V_{GS} (for small V_{DS}) for enhancement mode MOSFET. (b) Ideal $\sqrt{I_D}$ verses V_{GS} in saturation region for enhancement mode (curve A) and depletion mode (curve B) n-channel MOSFETs.

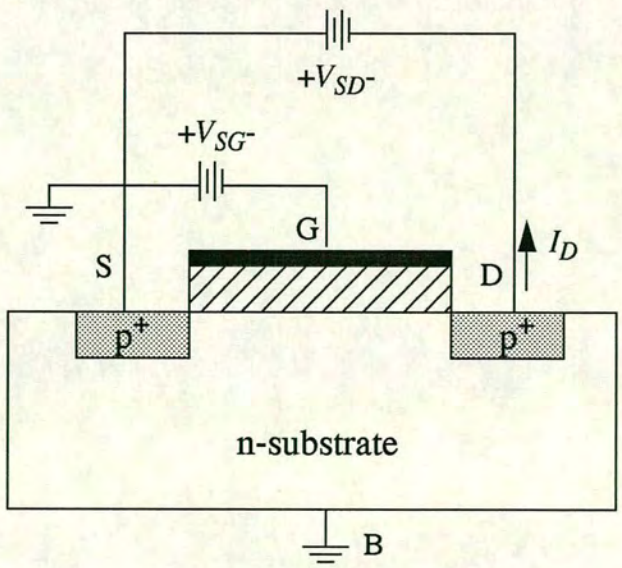


Figure 3.46 Cross section and bias configuration for a p-channel enhancement mode MOSFET.

Note the change in the sign in front of V_T and that the mobility is now the mobility is now the mobility of the holes in the hole inversion layer charge. The V_T for the p-channel enhancement mode MOSFET is negative and positive for the depletion mode p-channel device.

One assumption made in the derivation of the current-voltage relationship was that the charge neutrality condition given by Equation 3.43 was valid over the entire length of the channel, or implicitly that $Q_{SD}'(max)$ was constant. The space charge width varies between source and drain due to the drain-source voltage, it is widest when $V_{DS}>0$. A change in the space charge density along the channel length must be balanced by a corresponding change in the inversion layer charge. An increase in the space charge width means that the inversion layer charge is reduced, implying that the drain current and drain to source saturation voltage are less than the ideal values. The actual saturation drain current may be as much as 20 percent less than the predicted value due to this bulk charge effect.

3.3.4 Transconductance

The MOSFET transconductance is defined as the change in drain current with respect to the corresponding change in gate voltage,

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \qquad 3.63$$

The transconductance is sometimes referred to as the transistor gain.

For the n-channel MOSFET operating in the nonsaturation region, using Equation 3.55,

$$g_{mL} = \frac{\partial I_D}{\partial V_{GS}} = \frac{W\mu_n C_{ox}}{L} V_{DS} \qquad 3.64$$

The transconductance increases linearly with V_{DS} but is independent of V_{GS} in the nonsaturation region.

The I-V characteristics of an n-channel MOSFET in the saturation region were given by Equation 3.58. The transconductance in this region of operation is given by

$$g_{mL} = \frac{\partial I_D}{\partial V_{GS}} = \frac{W\mu_n C_{ox}}{L}(V_{GS} - V_T) \qquad 3.65$$

In the saturation region, the transconductance is a function of the geometry of the device as well as carrier mobility and threshold voltage. The transconductance increases as the width of the device increases, and as the channel length and oxide thickness decrease.

3.3.5 Substrate Bias Effects

In all of the analysis so far, the substrate, or body, has been connected to the source and held at ground potential. Figure 3.47a shows an n-channel MOSFET and the associated double subscripted voltage variables. The source to substrate pn junction must always be zero or reverse biased so that V_{SB} must always be greater than or equal to zero. If V_{SB} =0, threshold is defined as the condition when $\phi_s = 2\phi_{fp}$, and is shown in Figure 3.47b. However, these electrons are at a higher potential energy than are the electrons in the source. The newly created electrons will move laterally and flow out of the source terminal. When $\phi_s = 2\phi_{fp} + V_{SB}$, the surface reaches an equilibrium inversion condition. The energy-band diagram for this condition is shown in Figure 3.47c. The curve represented as E_{Fn} is the Fermi level from the p-substrate through the reverse-biased source-substrate junction to the source contact.

The space charge region width under the oxide increases from the original x_{dT} value when a reverse-biased source-substrate junction voltage is applied. With an applied V_{SB} >0, there is more charge associated with this region. Considering the charge neutrality condition through the MOS structure, the positive charge on the top metal gate must increase to compensate for the increased negative space charge in order to reach the threshold inversion point. So when V_{SB} >0, the threshold voltage of the n-channel MOSFET increases. When V_{SB} =0,

$$Q_{SD}'(max) = -eN_a x_{dT} = -\sqrt{2e\varepsilon_s N_a(2\phi_{fp})} \qquad 3.66$$

When V_{SB} >0, the space charge width increases to give

$$Q_{SD}' = -eN_a x_d = -\sqrt{2e\varepsilon_s N_a(2\phi_{fp} + V_{SB})} \qquad 3.67$$

The change in the space charge density is then

$$\Delta Q_{SD}' = -\sqrt{2e\varepsilon_s N_a}[\sqrt{2\phi_{fp} + V_{SB}} - \sqrt{2\phi_{fp}}] \qquad 3.68$$

To reach the threshold condition, the applied gate voltage must be increased. The change in threshold voltage can be written as

$$\Delta V_T = -\frac{\Delta Q_{SD}'}{C_{ox}} = \frac{\sqrt{2e\varepsilon_s N_a}}{C_{ox}}[\sqrt{2\phi_{fp} + V_{SB}} - \sqrt{2\phi_{fp}}] \qquad 3.69$$

where $\Delta V_T = V_T(V_{SB}>0) - V_T(V_{SB}=0)$. V_{SB} must always be positive for the n-channel device so that ΔV_T is positive. The threshold voltage will therefore increase as a function of source-substrate junction voltage for the n-channel MOSFET.

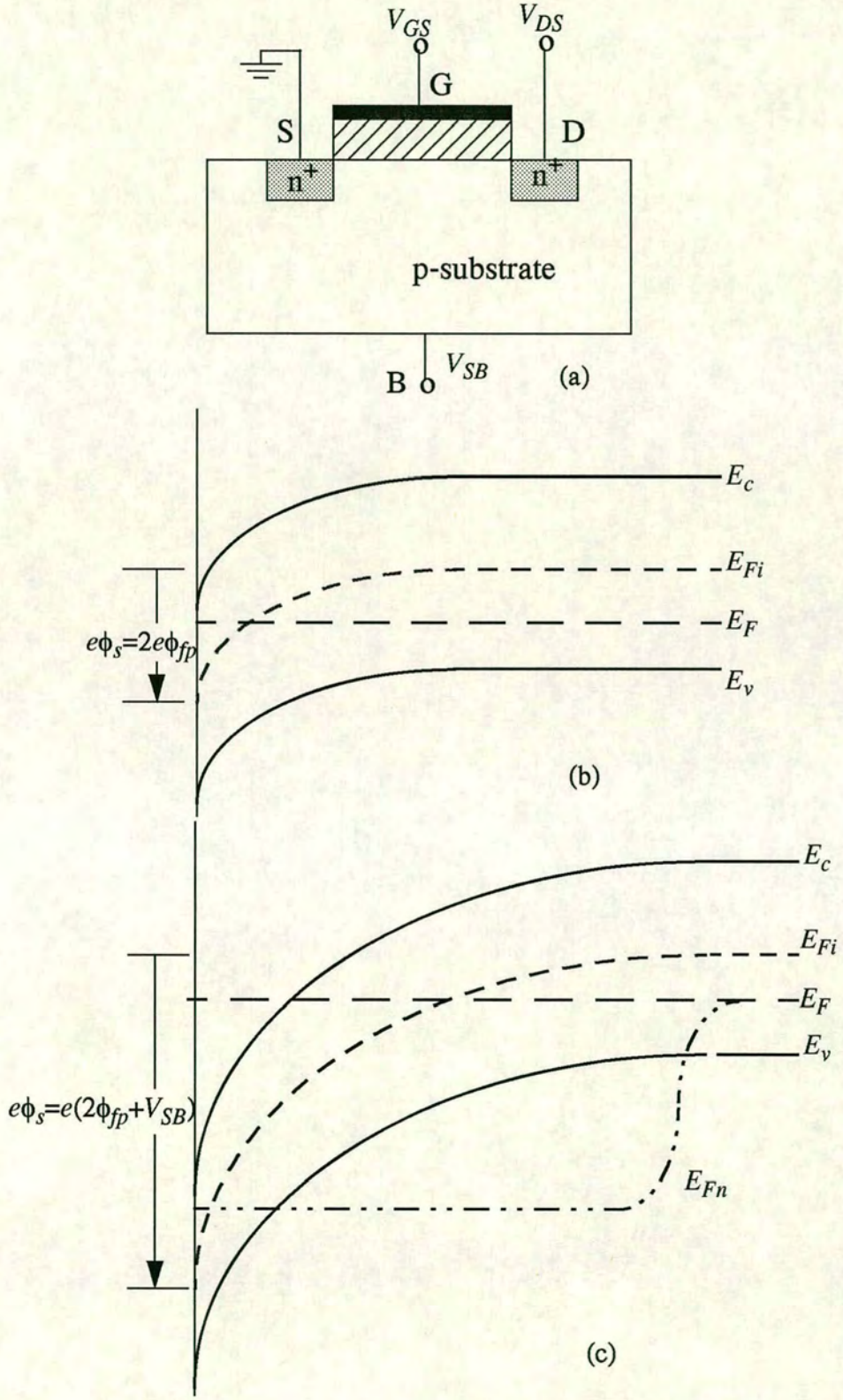


Figure 3.47 (a) Applied voltages on an n-channel MOSFET. (b) Energy-band diagram at inversion point when V_{SB}=0. (c) Energy-band diagram at inversion point when V_{SB}>0 is applied.

3.4 Non-Ideal Effects [33,34]

In this section, four effects will be considered that cause deviations from the ideal derivations; subthreshold conduction, channel length modulation, mobility variations and velocity saturation.

3.4.1 Subthreshold Conduction

The ideal current-voltage relationship predicts zero drain current when the gate to source voltage is less than or equal to the threshold voltage. Experimentally, I_D, is not zero when $V_{GS} \leq V_T$. Figure 3.48 shows a comparison between the ideal characteristics that were previously derived and experimentally derived results. The drain current, which exists for $V_{GS} \leq V_T$, is known as the subthreshold current.

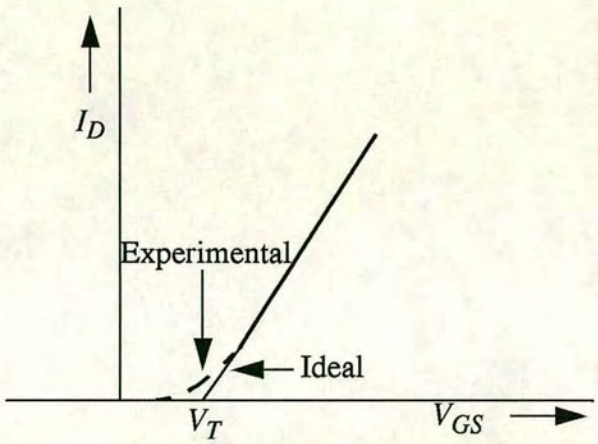


Figure 3.48 Comparison of ideal and experimental plots of $\sqrt{I_D}$ verses V_{GS}.

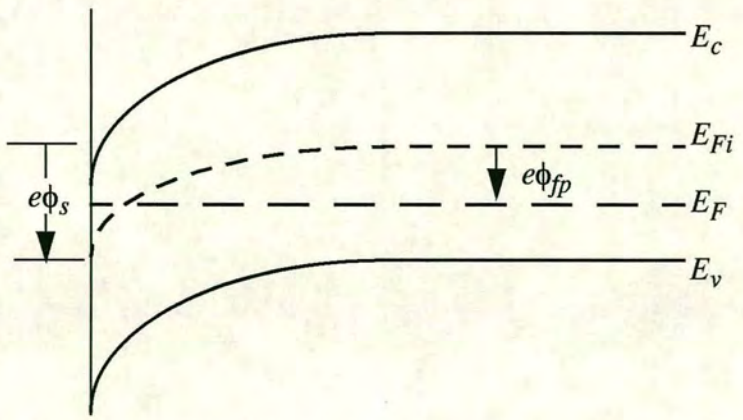


Figure 3.49 Energy-band diagram when $\phi_{fp}<\phi_s<2\phi_{fp}$.

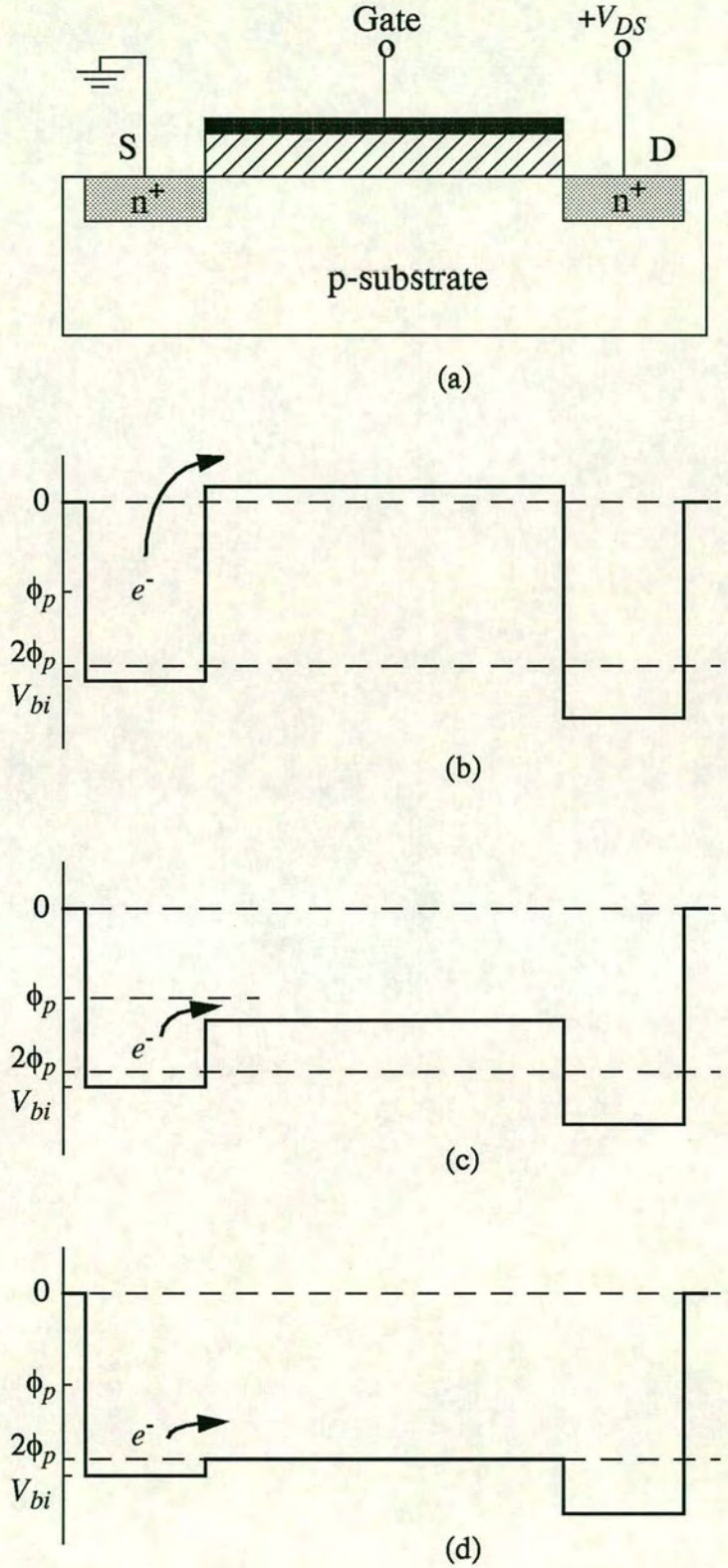


Figure 3.50 (a) Cross section along channel length of n-channel MOSFET. Potential diagrams along channel length at (b) accumulation, (c) weak inversion, and (d) inversion.

The energy-band diagram of a MOS structure with a p-type substrate biased so that $\phi_s<2\phi_{fp}$ is shown in Figure 3.49. The Fermi level is closer to the conduction band than the valence band, so the semiconductor surface develops the characteristics of a lightly doped n-type material. Some conduction between the n^+ source and drain contacts through this weakly inverted channel would therefore be observed. The condition for $\phi_{fp}<\phi_s<2\phi_{fp}$ is known as *weak inversion.*

Figure 3.50 shows the surface potential along the length of the channel at accumulation, weak inversion, and threshold for the case when a small drain voltage is applied. The bulk p-substrate is assumed to be at zero potential. Figures 3.50b and 3.50c show that for the accumulation and weak inversion cases, there is a potential barrier between the n^+ source and channel region which the electrons must overcome in order to generate a channel current. The channel current is an exponential function of V_{GS}. In the inversion mode, shown in Figure 3.50d, the barrier is so small that we loose the exponential dependence since the junction is now more like an ohmic contact. If V_{DS} is larger than a few *(kT/e)* volts, then the subthreshold current is independent of V_{DS}.

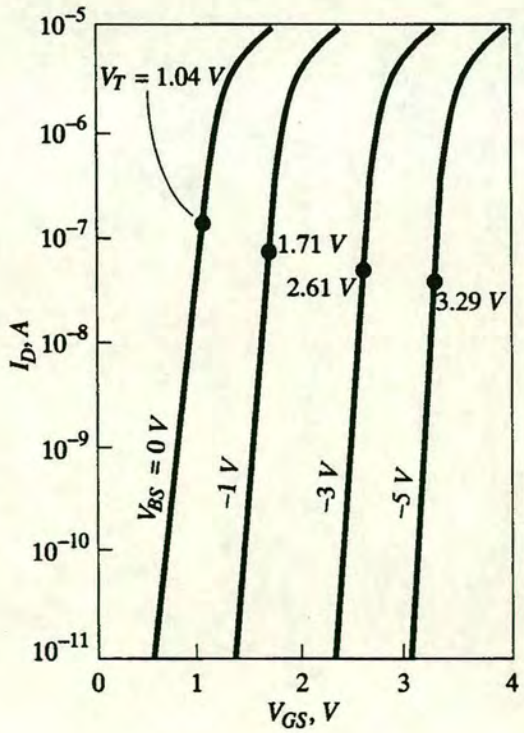


Figure 3.51 Subthreshold current-voltage characteristics for several values of substrate voltage (the threshold voltage is indicated on each curve) [35]

The subthreshold current is given by

$$I_D(sub) \propto \left[\exp\left(\frac{eV_{GS}}{kT}\right)\right] \cdot \left[1 - \exp\left(\frac{-eV_{DS}}{kT}\right)\right] \qquad 3.70$$

Figure 3.51 shows the exponential behaviour of the subthreshold current for several body to source voltages. Also shown on the curves are the threshold voltage values. Ideally, a change in gate voltage of approximately 60 mV produces an order of magnitude change in the subthreshold current. A detailed analysis of the subthreshold condition shows that the slope of the ln I_D verses V_{GS} curve is a function of the semiconductor doping and is also a function of the interface state density. The measurement of the slope of these curves has been used to experimentally determine the oxide-semiconductor interface state density.

If a MOSFET is biased at or even slightly below the threshold voltage, the drain current is not zero. The subthreshold current may add significantly to power dissipation in a large-scale integrated circuit in which millions of MOSFETs are used.

3.4.2 Channel Length Modulation

When a MOSFET is biased in the saturation region, the depletion region at the drain terminal extends laterally into the channel, reducing the effective channel length. Since the depletion region width is bias dependent, the effective channel length is also bias dependent and is modulated by the drain to source voltage. This channel length modulation effect is shown in Figure 3.52 for a n-channel MOSFET.

As a first approximation, the incremental change in the depletion layer at the drain is given by

$$\Delta L = \sqrt{\frac{2\varepsilon_s}{eN_a}}\left[\sqrt{\phi_{fp} + V_{DS}(sat) + \Delta V_{DS}} - \sqrt{\phi_{fp} + V_{DS}(sat)}\right] \qquad 3.71$$

where

$$\Delta V_{DS} = V_{DS} - V_{DS}(sat) \qquad 3.72$$

The applied drain to source voltage is V_{DS} and we are assuming that $V_{DS} > V_{DS}(sat)$. The drain terminal is very heavily doped so the depletion region extends only into the channel region.

Since the drain current is inversely proportional to the channel length,

$$I_D' = \left(\frac{L}{L - \Delta L}\right) I_D \qquad 3.73$$

where I_D' is the actual drain current and I_D is the ideal drain current. Since ΔL is a function of V_{DS} even though the transistor is biased in the saturation region. Figure 3.53 shows some typical I_D' verses V_{DS} curves with positive slopes in the saturation region due to channel length modulation. As the MOSFET dimensions become smaller, the change in the channel length ΔL becomes a larger fraction of the original length L and the channel length modulation becomes more severe.

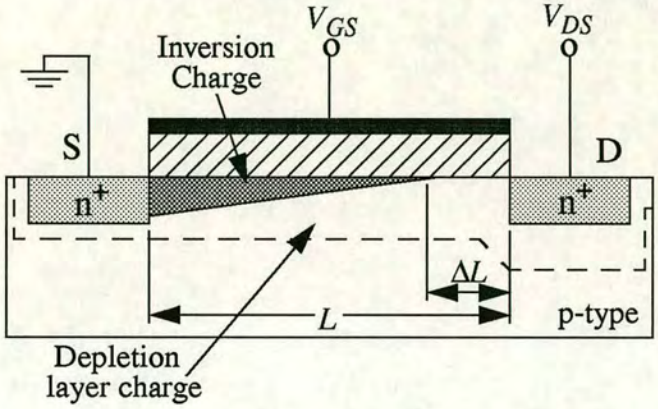


Figure 3.52 Cross section of n-channel MOSFET showing channel length modulation effect.

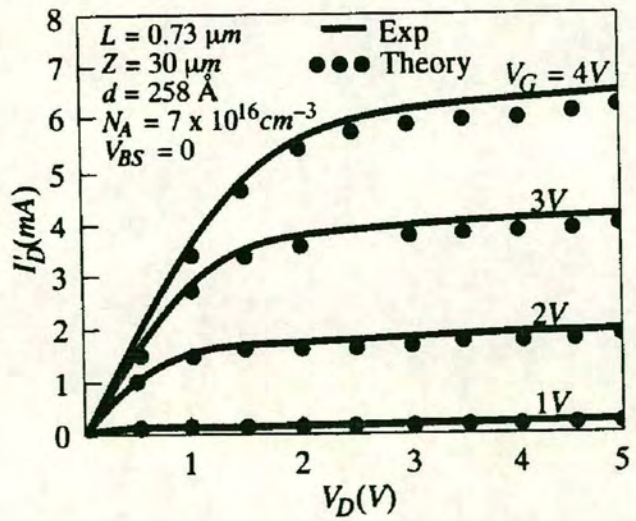


Figure 3.53 Current-voltage characteristics of a MOSFET showing short-channel effects [30].

3.4.3 Mobility Variation

In the derivation of the I-V relationship, the mobility was assumed to be constant, this assumption must be modified for two reasons. The first effect to be considered is the variation in mobility with gate voltage. The second reason for a mobility variation is that the effective carrier mobility decreases as the carrier approaches the velocity saturation limit. This effect will be discussed in the next section.

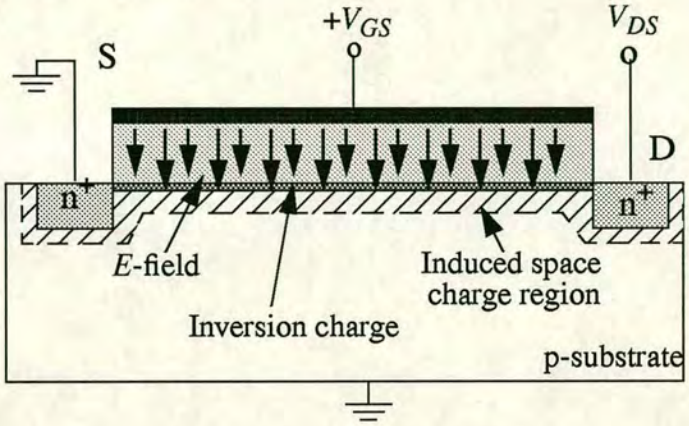


Figure 3.54 Vertical electric field in an n-channel MOSFET.

The inversion layer charge is induced by a vertical electric field, which is shown in Figure 3.54 for an n-channel device. A positive gate voltage produces a force on the electrons in the inversion layer toward the surface. As the electrons travel through the channel toward the drain, they are attracted to the surface, but then are repelled by localized Coulombic forces. This effect, schematically shown in Figure 3.55, is called *surface scattering*. The surface scattering effect reduces mobility. If there is a positive fixed oxide charge near the oxide-semiconductor interface, the mobility will be further reduced due to the additional coulomb interaction.

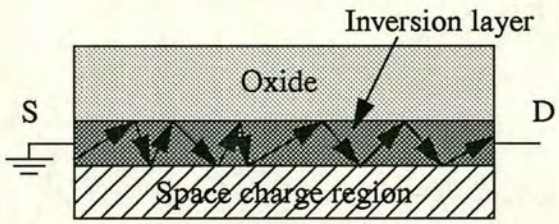


Figure 3.55 Schematic of carrier surface scattering effects.

The relationship between the inversion charge mobility and the transverse electric field is

measured experimentally. An effective transverse electric field can be defined as

$$E_{eff} = \frac{1}{\varepsilon_s}\left(|Q_{SD}'(max)| + \frac{1}{2}Q_n'\right) \qquad 3.74$$

The effective inversion charge mobility can be determined from the channel conductance as a function of gate voltage. Figure 3.56 shows the effective electron mobility at T=300°K for different doping levels and different oxide thicknesses. The effective mobility is only a function of the electric field at the inversion layer and is independent of oxide thickness. The effective mobility may be represented by

$$\mu_{eff} = \mu_0\left(\frac{E_{eff}}{E_0}\right)^{-1/3} \qquad 3.75$$

where μ_0 and E_0 are constants determined from experimental results.

The effective inversion charge mobility is a strong function of temperature because of lattice scattering. As the temperature is reduced, the mobility increases.

The effective mobility is a function of gate voltage through the inversion charge density in Equation 3.74. As the gate voltage increases, the carrier mobility decreases even further.

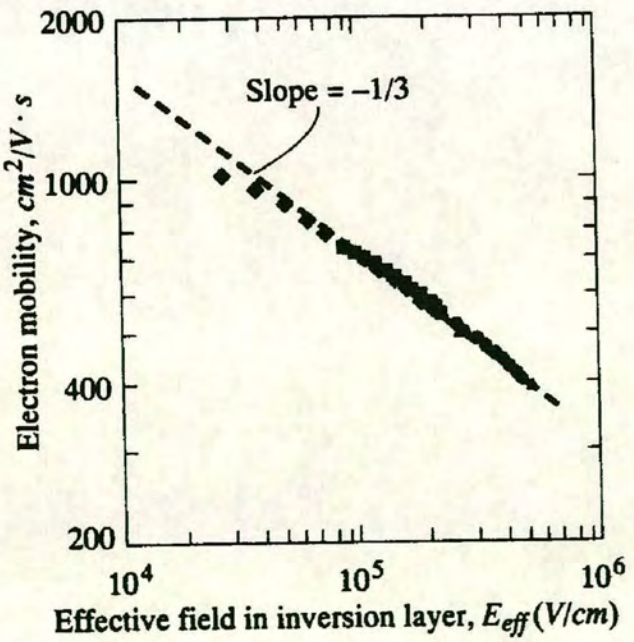


Figure 3.56 Measured inversion layer electron mobility verses electric field at the inversion layer [36]

3.4.4 Velocity Saturation

The mobility variation due to the horizontal drain to source electric field is caused by the velocity saturation effect. Velocity saturation is more prominent in shorter-channel devices where the horizontal electric field is larger.

In the ideal I-V relationship, current saturation occurred when the inversion charge density became zero at the drain terminal, or when

$$V_{DS} = V_{GS} - V_T \qquad 3.56$$

for the n-channel MOSFET. Velocity saturation can change this saturation condition. Velocity saturation will occur when the horizontal electric field is approximately 10^4 V/cm. If V_{DS}=5 volts in a device with a channel length of L=1μm, the average electric field is 5 x 10^4 V/cm. Velocity saturation, then is very likely to occur in short-channel devices.

The modified $I_D(sat)$ characteristics are described approximately by

$$I_D(sat) = WC_{ox}(V_{GS} - V_T)v_{sat} \qquad 3.76$$

where v_{sat} is the saturation velocity (approximately 10^7 cm/sec for electrons in bulk silicon) and C_{ox} is the gate oxide capacitance per cm^2. The saturation velocity will decrease slightly with applied gate voltage because of the vertical electric field and surface scattering. Velocity saturation will yield an $I_D(sat)$ value which is smaller than that predicted by the ideal relation and will yield a smaller $V_{DS}(sat)$ value than predicted. The $I_D(sat)$ current is also approximately linear with V_{GS} rather than having the ideal square law dependence predicted previously.

There are several models of mobility verses electric field. The most commonly used relation is

$$\mu = \frac{\mu_{eff}}{\left[1 + \left(\frac{\mu_{eff}E}{v_{sat}}\right)^2\right]^{1/2}} \qquad 3.77$$

Figure 3.57 shows a comparison of drain current verses drain to source voltage characteristics for constant mobility and for field-dependent mobility. The smaller values of $I_D(sat)$ and the

approximate linear dependence on V_{GS} may be noted for the field-dependent mobility curves. The transconductance is found from

$$g_{mL} = \frac{\partial}{\partial V_{GS}} I_D(sat) = WC_{ox}v_{sat} \qquad 3.78$$

which is now independent of $V_{GS}$and V_{DS} when velocity saturation occurs. The drain current is saturated by the velocity saturation effect, which leads to a constant transconductance.

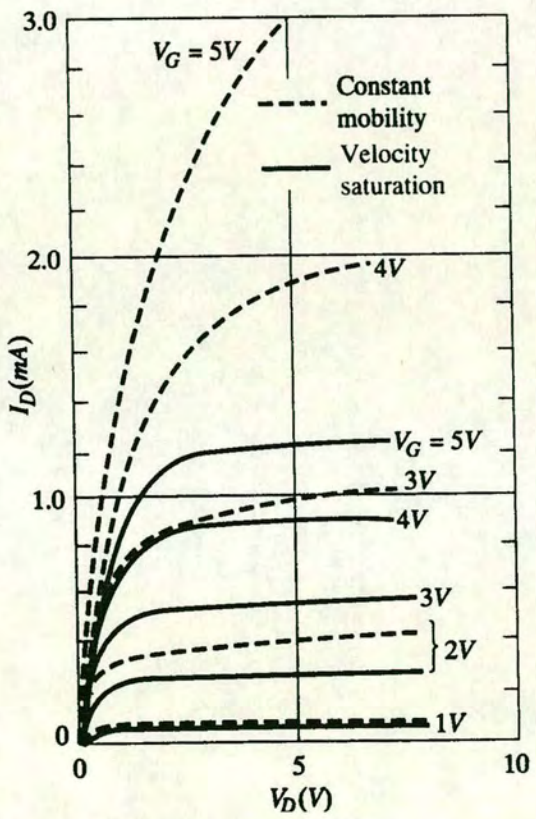


Figure 3.57 Comparison of I_D verses V_D characteristics for constant mobility (dashed curves) and for field-dependent mobility and velocity saturation effects (solid curves) [30].

When velocity saturation occurs, the cutoff frequency is given by,

$$f_T = \frac{g_m}{2\pi C_G} = \frac{WC_{ox}v_{sat}}{2\pi(C_{ox}WL)} = \frac{v_{sat}}{2\pi L} \qquad 3.79$$

where the parasitic capacitances are assumed to be negligible.

3.4.5 Breakdown Voltages

Several voltage breakdown mechanisms will be considered for the MOSFET. The topic of voltage breakdown across the oxide will however be covered in Chapter 4.

Avalanche Breakdown. Avalanche breakdown can occur by impact ionization in the space charge region near the drain terminal. In an ideal planar one-sided pn junction, breakdown is a function primarily of the doping concentration in the low-doped region of the junction. For the MOSFET, the low-doped region corresponds to the semiconductor substrate. If the p-type substrate doping is N_a=3 x 10^{16} cm^{-3}, for example, the pn junction breakdown voltage would be approximately 25 volts for a planar junction. However, the n^+ drain contact may be a fairly shallow diffused region with a large curvature. The electric field in the depletion region tends to be concentrated at the curvature, which lowers the breakdown voltage. This curvature effect is shown in Figure 3.58.

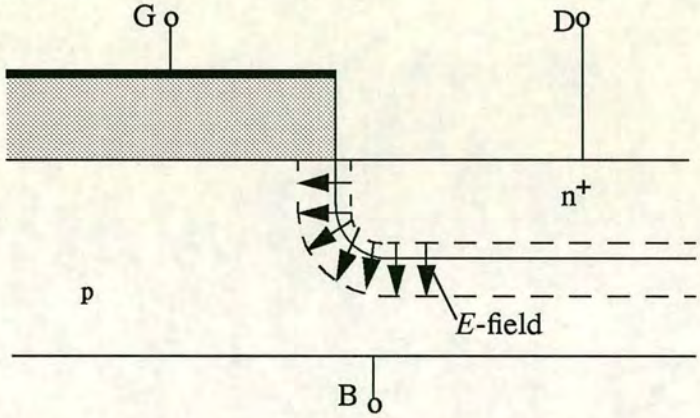


Figure 3.58 Curvature effect on the electric field in the junction.

Near Avalanche and Snapback Breakdown. Another breakdown mechanism results in the S-shaped breakdown curve shown in Figure 3.59.

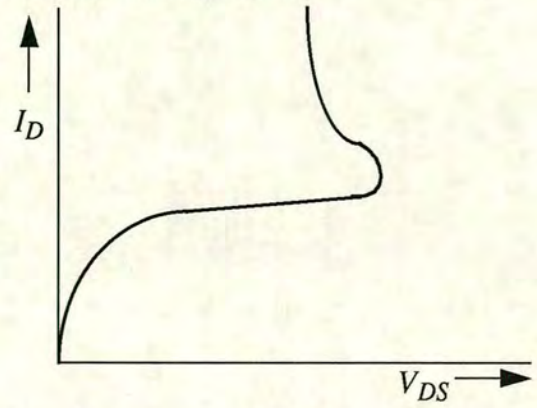


Figure 3.59 Current-voltage characteristic showing the snapback breakdown effect.

This breakdown process is due to second order effects and can be explained with the aid of Figure 3.60. The n-channel enhancement mode MOSFET geometry in Figure 3.60a shows the n-type source and drain contacts along with the p-type substrate. The source and body are at ground potential. The n(source)-p(substrate)-n(drain) structure also forms a parasitic bipolar transistor. The equivalent circuit is shown in Figure 3.60b.

Figure 3.61a shows the device when avalanche breakdown is just beginning in the space charge region near the drain. The avalanche breakdown does not just suddenly occur at a particular voltage, but is a gradual process that starts at low current levels and for electric fields below the breakdown field. The electrons generated by the avalanche process flow into the drain and contribute to the substrate to the body terminal. Since the substrate has a non-zero resistance, a voltage drop is produced as shown. This potential difference drives the source-to-substrate pn junction into forward bias near the source terminal. The source is heavily doped n-type thus, a large number of electrons can be injected from the source contact into the substrate under forward bias. This process will become severe as the voltage drop in the substrate approaches 0.6 to 0.7 volt. A fraction of the injected electrons will diffuse across the parasitic base region into the reverse-biased drain space charge region where they also add to the drain current.

The avalanche breakdown process is a function of not only the electric field but also the number of carriers involved. The rate of avalanche breakdown increases as the number of carriers in the drain space charge region increases. There is now a regenerative or positive feedback mechanism. Avalanche breakdown near the drain terminal produces the substrate current, which produces the forward-biased source-substrate pn junction voltage. The forward-biased junction injects carriers that can diffuse back to the drain and increase the avalanche process. The positive feedback produces an unstable system.

The snapback or negative resistance portion of the curve shown in Figure 3.59 can now be explained using the parasitic bipolar transistor. The potential of the base of the bipolar transistor near the emitter (source) is almost floating since this voltage is determined primarily by the avalanche-generated substrate current rather than an externally applied voltage.

For the open-base bipolar transistor shown in Figure 3.61b,

$$I_C = \alpha I_E + I_{CB0} \qquad 3.80$$

where α is the common base current gain and I_{CB0} is the base-collector leakage current.

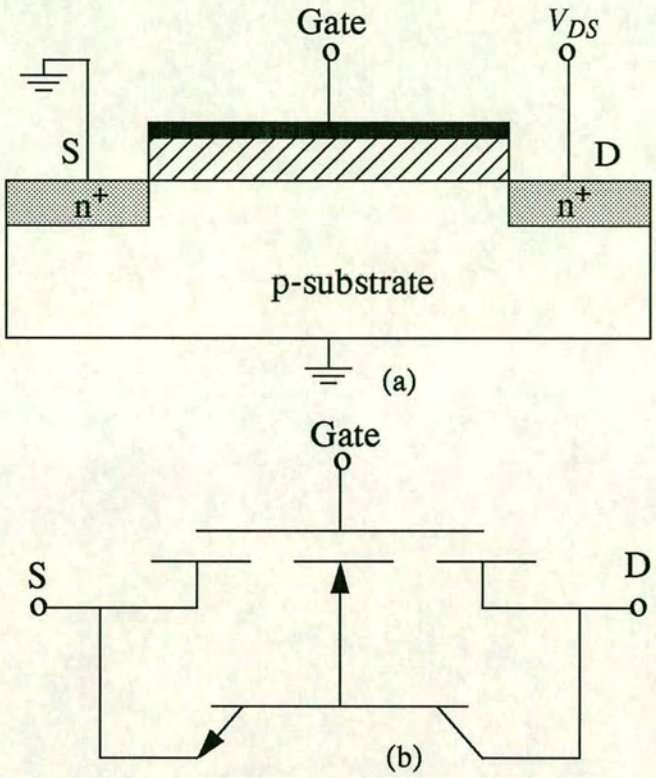


Figure 3.60 (a) Cross section of the n-channel MOSFET. (b) Equivalent circuit including the parasitic bipolar transistor.

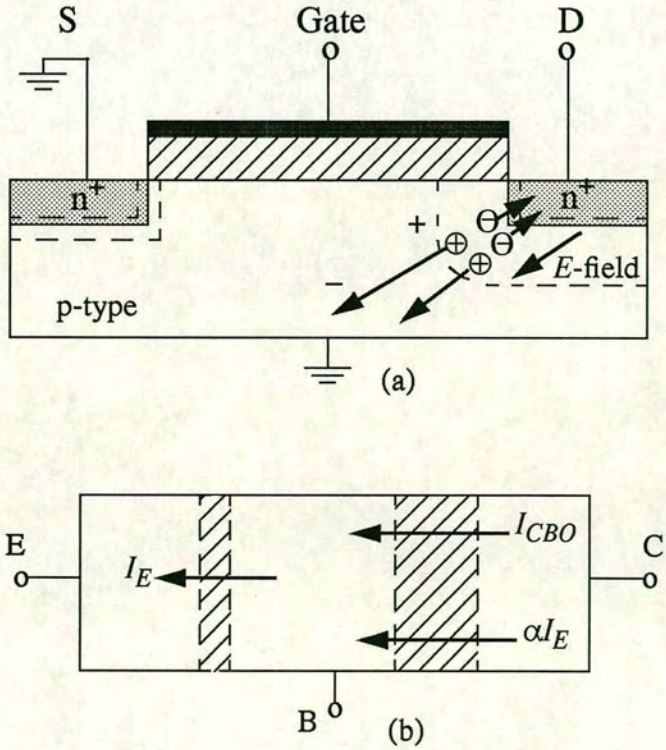


Figure 3.61 (a) Substrate current-induced voltage drop caused by avalanche multiplication at the drain. (b) Currents in the parasitic bipolar transistor.

For an open-base, $I_C = I_E$, so Equation 3.80 becomes

$$I_C = \alpha I_C + I_{CB0} \qquad 3.81$$

At breakdown, the current in the B-C junction is multiplied by the multiplication factor M,

$$I_C = M(\alpha I_C + I_{CB0}) \qquad 3.82$$

Solving for I_C,

$$I_C = \frac{M I_{CB0}}{1 - \alpha M} \qquad 3.83$$

Breakdown is defined as the condition which produces $I_C \rightarrow \infty$. For a single reverse-biased pn junction, $M \rightarrow \infty$ at breakdown. From Equation 3.83 however, breakdown is defined to be the condition when $\alpha M \rightarrow 1$ or, for the open-base condition, breakdown occurs when $M \rightarrow 1/\alpha$, which is a much lower multiplication factor for the simple pn junction.
An empirical relation for the multiplication factor is usually written as

$$M = \frac{1}{1 - (V_{CE}/V_{BD})^m} \qquad 3.84$$

where m is an empirical constant in the range of 3 to 6 and V_{BD} is the junction breakdown voltage.

The common base current gain factor α is a strong function of collector current for small values of collector current. At low currents, the recombination current in the B-E junction is a significant fraction of the total current so that the common base current gain is small. As the collector current increases, the value of α increases. As avalanche breakdown begins and I_C is small, particular values of M and V_{CE} are required to produce the condition of αM =1. As the collector current increases, α increases; therefore smaller values of M and V_{CE} are required to produce the avalanche breakdown condition. The snapback, or negative resistance, breakdown characteristic is then produced.

Only a fraction of the injected electrons from the forward-biased source-substrate junction

are collected by the drain terminal. A more exact calculation of the snapback characteristic would take this into account.

The snapback effect can be minimized by using a heavily doped substrate that will prevent any significant voltage drop from being developed. A thin epitaxial p-type layer with the proper doping concentration to produce the required threshold voltage can be grown on a heavily doped substrate.

Near Punch-Through Effects. Punch-through is the condition at which the drain-to-source space charge extends completely across the channel region to the source-to-substrate space charge region. In this situation, the barrier between the source and drain is completely eliminated and a very large drain current would exist. The drain current will however, begin to increase rapidly before the actual punch-through condition is reached. This characteristic is referred to as the near punch-through condition. Figure 3.62a shows the ideal energy-band diagram from source to drain for a long n-channel MOSFET for the case when $V_{GS}<V_T$ and when the drain-to-source voltage is relatively small. The large potential barriers prevent significant current between the drain and source. Figure 3.62b shoes the energy-band diagram when a relatively large drain voltage V_{DS2} is applied. The space charge region near the drain terminal is beginning to interact with the source space charge region and the potential barrier is being lowered. Since the current is an exponential function of barrier height, the current will increase very rapidly with drain voltage once this near punch-through condition has been reached. Figure 3.63 shows some typical characteristics of a short-channel device with a near punch-through condition.

For a channel doping of $10^{16}cm^{-3}$ and source/drain dopings of $10^{19}cm^{-3}$, the two space charge regions will begin to interact when the abrupt depletion layers are approximately $0.25\mu m$ apart. The drain voltage at which this near punch-through condition, also known as drain-induced barrier lowering, occurs at a significantly lower punch-through voltage than for a long-channel device.

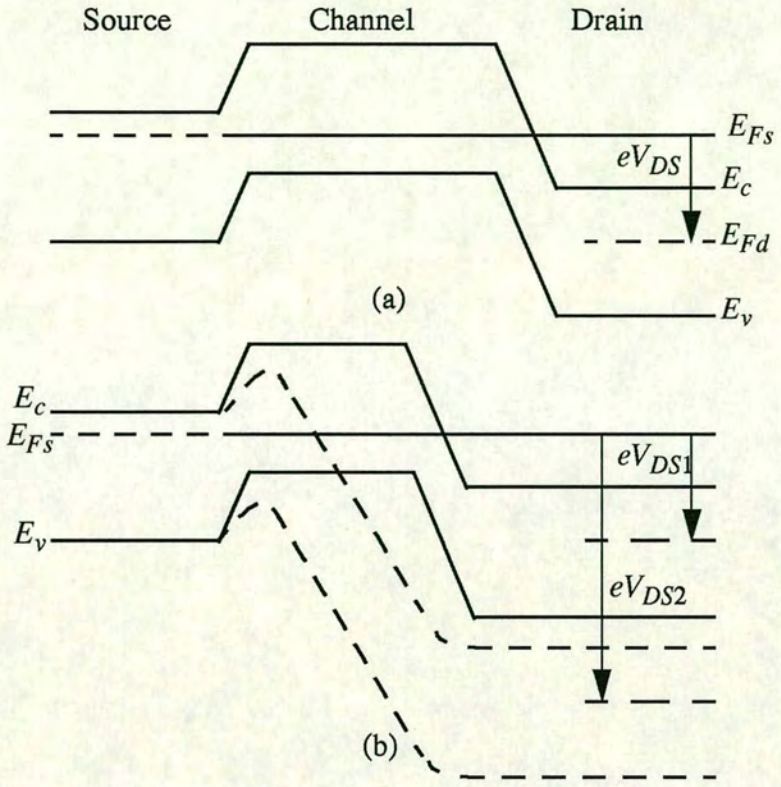


Figure 3.62 (a) Equipotential plot along the surface of a long-channel MOSFET. (b) Equipotential plot along the surface of a short-channel MOSFET before and after punch-through.

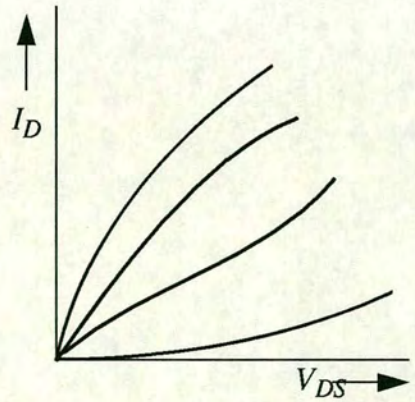


Figure 3.63 Typical I-V characteristics of a MOSFET exhibiting punch-through effects.

3.5 Small Device Geometries

3.5.1 Short-Channel Effects

In the ideal MOSFET, the threshold voltage was derived using the concept of charge neutrality in which the sum of charges in the metal, oxide, inversion layer, and semiconductor space charge region is zero. The gate area was also assumed to be the same as the active area in the semiconductor. Using this assumption, only equivalent surface charge densities were considered and the effects on threshold voltage from the source and drain space charge regions extending into the active channel were neglected.

Figure 3.64a shows the cross section of a long n-channel MOSFET at flat-band, with zero source and drain voltage applied. The space charge regions at the source and drain extend into the channel region, but occupy only a small fraction of the entire channel region. The gate voltage, then will control essentially all of the space charge induced in the channel region at inversion as shown in Figure 3.64b

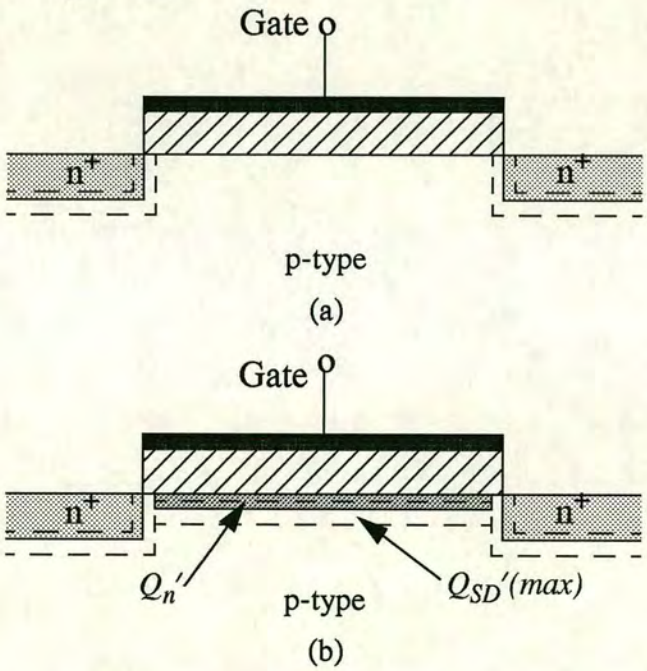


Figure 3.64 Cross section of a long n-channel MOSFET (a) at flat-band and (b) at inversion.

As the channel length decreases, the fraction of charge in the channel region controlled by the gate decreases. This effect can be seen in Figure 3.65 for the flat-band condition. As the drain voltage increases, the reverse-biased space charge region at the drain extends further into the channel area and the gate will control even less bulk charge. The amount of charge in the channel region, $Q_{SD}'(max)$, controlled by the gate, affects the threshold voltage as can be seen

from Equation 3.85,

$$V_{TN} = (|Q_{SD}'(max)| - Q_{SS}')\left(\frac{t_{ox}}{\varepsilon_{ox}}\right) + \phi_{ms} + 2\phi_{fp} \qquad 3.85$$

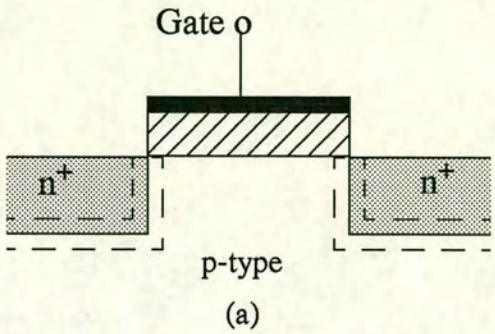


Figure 3.65 Cross section of a short n-channel MOSFET at flat-band.

The affect on the threshold voltage from short-channel effects can be quantitatively determined by considering the parameters shown in Figure 3.66 The source and drain junctions are characterized by a diffused junction depth r_j. The lateral diffusion distance under the gate is assumed to be the same as the vertical diffusion distance. The source, drain and body contacts are initially all at ground potential.

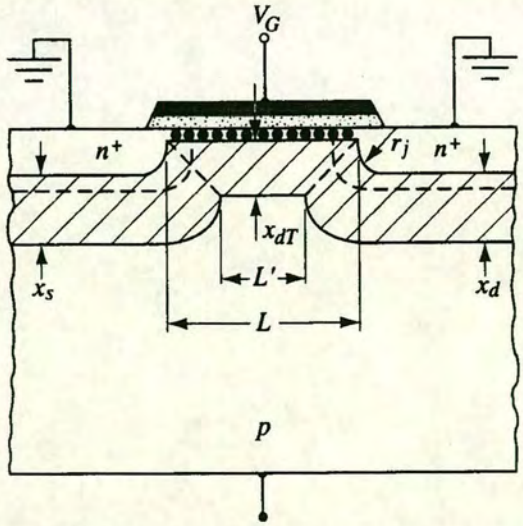


Figure 3.66 Charge sharing in the short-channel threshold voltage model [37]

The basic assumption in this analysis is that the bulk charge in the trapezoidal region under the gate is controlled by the gate. The potential difference across the bulk space charge region is $2\phi_{fp}$ at the threshold inversion point and the built-in potential barrier height of the source and drain junctions is also approximately $2\phi_{fp}$, implying that the three space charge widths are

essentially equal. Then

$$x_s \approx x_d \approx x_{dT} \qquad 3.86$$

Using the geometrical approximation, the average bulk charge per unit area Q_B' in the trapezoid is

$$|Q_B'| \cdot L = eN_a x_{dT}\left(\frac{L+L'}{2}\right) \qquad 3.87$$

From the geometry,

$$\frac{L+L'}{2L} = \left\{1 - \frac{r_j}{L}\left[\sqrt{1+\frac{2x_{dT}}{r_j}} - 1\right]\right\} \qquad 3.88$$

Then

$$|Q_B'| = eN_a x_{dT}\left\{1 - \frac{r_j}{L}\left[\sqrt{1+\frac{2x_{dT}}{r_j}} - 1\right]\right\} \qquad 3.89$$

Equation 3.89 is now used in place of $|Q_{SD}'(max)|$ in the expression for the threshold voltage. Since $|Q_{SD}'(max)| = eN_a x_{dT}$,

$$\Delta V_T = -\frac{eN_a x_{dT}}{C_{ox}}\left\{\frac{r_j}{L}\left[\sqrt{1+\frac{2x_{dT}}{r_j}} - 1\right]\right\} \qquad 3.90$$

where

$$\Delta V_T = V_{T(shortchannel)} - V_{T(longchannel)} \qquad 3.91$$

As the channel length decreases, the threshold voltage shifts in the negative direction so that an n-channel MOSFET shifts toward depletion mode. The effect of short channels becomes more pronounced as the channel length is reduced further.

The shift in threshold voltage with channel length for an n-channel MOSFET is shown in Figure 3.67. As the substrate doping increases, the initial threshold voltage increases and the short-channel threshold shift also becomes larger. The short-channel effects on threshold voltage do not become significant until the channel length becomes less than approximately

2μm. The threshold voltage shift also becomes smaller as the diffusion depth r_j becomes smaller so that very shallow junctions reduce the threshold voltage dependence on channel length. Equation 3.90 was derived using the assumption that the source, channel and drain space charge widths were all equal. If a voltage is applied to the drain, the space charge width at the drain terminal widens, which makes L' smaller, and the amount of bulk charge controlled by the gate voltage decreases. This effect makes the threshold voltage a function of drain voltage. As the drain voltage increases, the threshold voltage of an n-channel MOSFET decreases. The threshold voltage verses channel length is plotted in Figure 3.68 for two values of drain-to-source voltage and two values of body-to-source voltage.

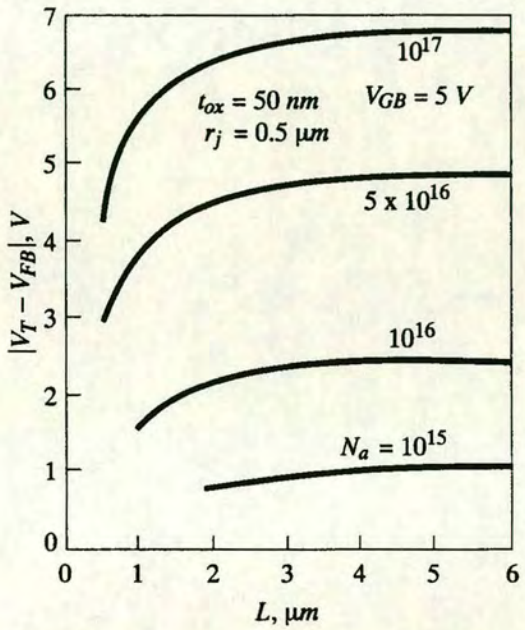


Figure 3.67 Threshold voltage versus channel length for various substrate dopings [37].

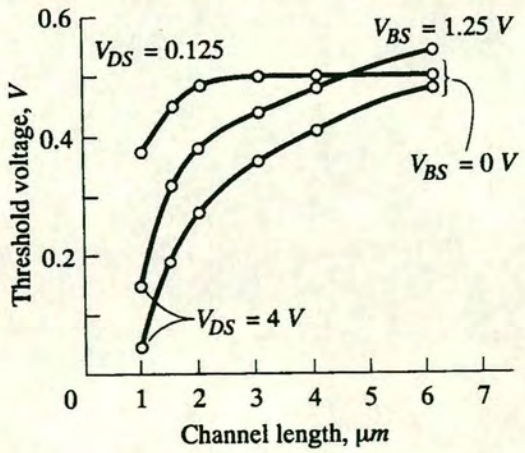


Figure 3.68 Threshold voltage versus channel length for two values of drain-to-source and body-to-source voltage [36].

3.5.2 Narrow-Channel Effects

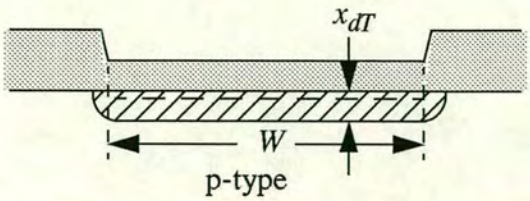


Figure 3.69 Cross section of a n-channel MOSFET showing the depletion region along the width of the device.

Figure 3.69 shows the cross section along the channel width of an n-channel MOSFET biased at inversion. The current is perpendicular to the channel width through the inversion charge. Note in the figure that there is an additional space charge region at each end of the channel width. This additional charge is controlled by the gate voltage but was not included in the derivation of the ideal threshold voltage relation. The threshold voltage expression must be modified to include this additional charge.

Neglecting short-channel effects, the gate controlled bulk charge can be written as,

$$Q_B = Q_{B0} + \Delta Q_B \qquad 3.92$$

where Q_B is the total bulk charge, Q_{B0} is the ideal bulk charge and ΔQ_B is the additional bulk charge at the ends of the channel width. For a uniformly doped p-type semiconductor biased at the threshold inversion point,

$$|Q_{B0}| = eN_a WLx_{dT} \qquad 3.93$$

and

$$\Delta Q_B = eN_a Lx_{dT}(\xi x_{dT}) \qquad 3.94$$

where ξ is a fitting parameter that accounts for the lateral space charge width. The lateral space charge width may not be the same as the vertical width x_{dT} due to the thicker field oxide

at the ends, and/or due to the nonuniform semiconductor doping created by an ion implantation. If the ends were a semicircle, then $\xi = \pi/2$. Now,

$$|Q_B| = |Q_{B0}| + |\Delta Q_B| = eN_a WLx_{dT}\left(1 + \frac{\xi x_{dT}}{W}\right) \qquad 3.95$$

The effect of the end space charge regions becomes significant as the width W decreases and the factor (ξx_{dT}) becomes a significant fraction of the width W.

The change in threshold voltage due to the additional space charge is

$$\Delta V_T = \frac{eN_a x_{dT}}{C_{ox}}\left(\frac{\xi x_{dT}}{W}\right) \qquad 3.96$$

The shift in threshold due to a narrow channel is in the positive direction for the n-channel MOSFET. As the width W becomes smaller, the shift in threshold voltage becomes larger.

Figure 3.70 shows the threshold voltage as a function of channel width. The threshold shift begins to become apparent for channel widths that are large compared to the induced space charge width.

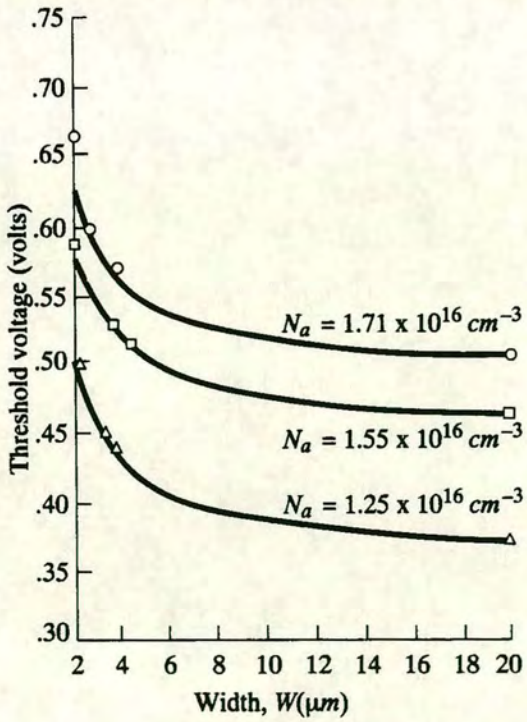


Figure 3.70 Threshold voltage versus channel width (solid curves, theoretical; points, experimental) [38].

Figures 3.71a and 3.71b show qualitatively the threshold voltage shifts due to short-channel and narrow-channel effects, respectively, in n-channel MOSFETs. The narrow-channel device produces a larger threshold voltage and the short-channel device produces a smaller threshold voltage. For devices exhibiting both short-channel and narrow-channel effects, the two models need to be combined into a three-dimensional volume approximation of the space charge region controlled by the gate.

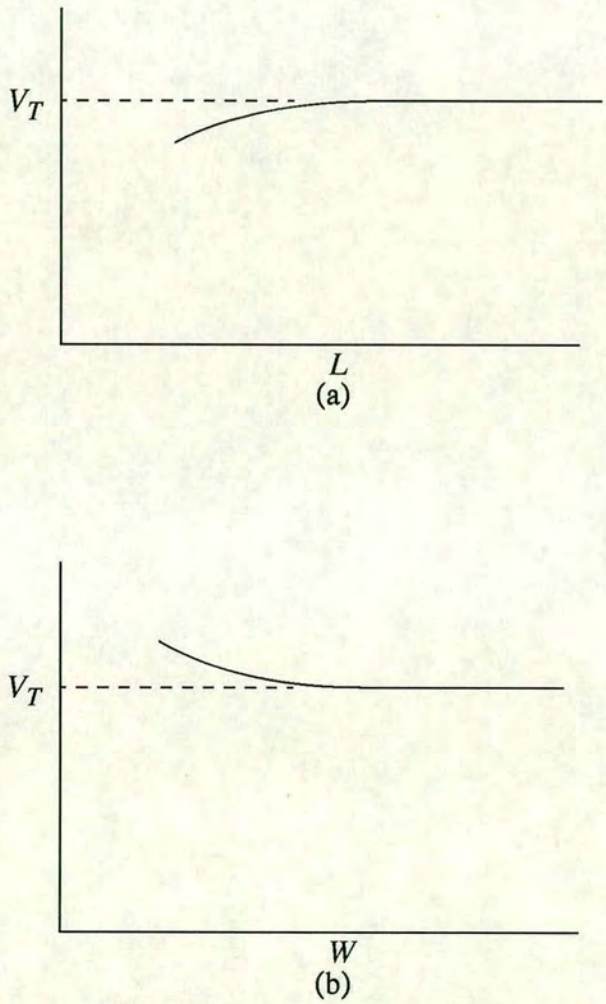


Figure 3.71 Qualitative variation of threshold voltage (a) with channel length and (b) with channel width.

CHAPTER 4

Reliability of Thin Gate Oxides

The reliability of thin gate oxide is described in terms of test methods and failure mechanisms. This understanding is required to assess the reliability of oxides grown by the nitrogen implanted silicon method since this is a key element to the success of this method for growing different thicknesses of gate oxide on a single circuit.

4.1 Carrier injection into the Oxide

Carriers injected into the gate oxide from the silicon substrate can cause either catastrophic damage to the oxide or can change the device operating characteristics. The injection of carriers through the energy barrier is associated with 'cold' carriers and is accomplished by either Fowler-Nordheim or Direct tunnelling. Carriers which are able to surmount the energy barrier are called 'hot' carriers because they possess kinetic energies higher than the equilibrium temperature of the silicon lattice at the point of injection. The topic of hot carrier injection into the oxide will be covered in Chapter 5.

4.1.1 Fowler-Nordheim Tunnelling

Fowler-Nordheim (FN) tunnelling [39] describes the injection of electrons into the conduction band in the oxide through the triangular part of the energy barrier. Once inside the oxide, the electrons are accelerated by the oxide field to the anode (gate) and produce a gate current. Figure 4.1 shows the energy-band diagram for this case where a positive voltage is applied to the gate. Also shown in the figure is the possibility of trap assisted FN tunnelling.

An expression has been derived for the FN tunnelling current density J,

where $A_F = 1.25 \times 10^{-6}$ A/V^2, $B \sim 240$MV/cm and ε_{ox} is the oxide field in V/cm. As the electric field over the oxide is increased, there is a higher probability that an electron will tunnel through the oxide and a higher leakage current is generated.

4.1.2 Direct Tunnelling

As the oxide is thinned, there is a higher probability that an electron will tunnel straight through the trapezoid-shaped energy barrier. This tunnelling behaviour is known to occur in oxides thinner than 6 nm at low voltages. As with FN tunnelling, it has also been postulated that traps located in the forbidden gap could assist the tunnelling mechanism. The

mechanisms of direct tunnelling and trap assisted direct tunnelling are shown in Figure 4.2

For oxides thinner than 3 nm, the direct tunnelling current becomes so large that carriers are removed from the silicon faster than they are thermally generated so that an inversion layer is prevented from forming. This sets the limit to the scaling of oxides to ~ 3 nm thickness.

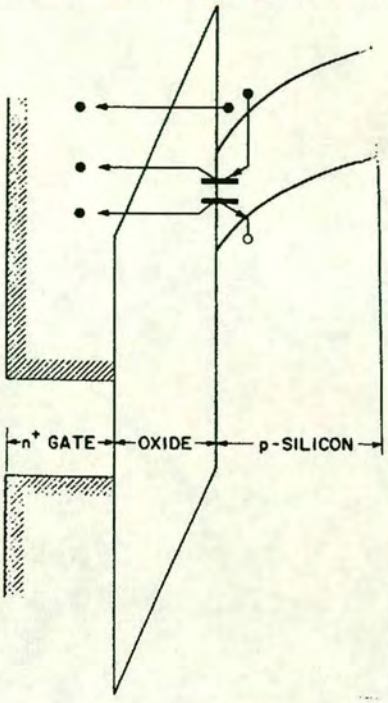


Figure 4.1 Energy-band diagram for Fowler-Nordheim tunnelling. Also shown are interface-trap assisted FN tunnelling mechanisms with interface traps identified by thick horizontal bars [40].

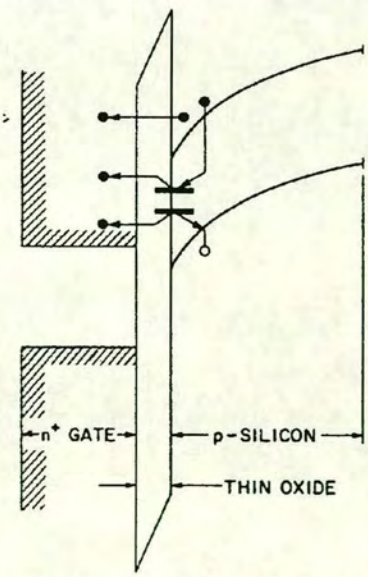


Figure 4.2 Energy-band diagram for Direct tunnelling through the oxide for thin oxides. Also shown are interface-trap assisted Direct tunnelling mechanisms with interface traps identified by thick horizontal bars [40].

Figure 4.3 shows the experimentally measured current density and FN tunnelling prediction for thin oxides with increasing applied gate voltage. The observed discrepancies at low gate voltages are attributed to the direct tunnelling current density

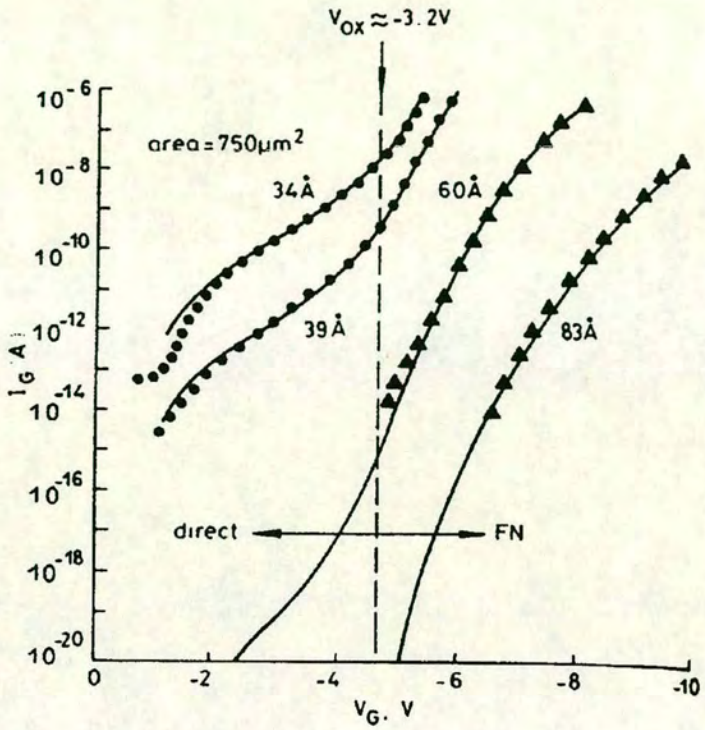


Figure 4.3 Comparison of measured oxide conduction with predicted FN tunnelling current at increasing gate voltages [40].

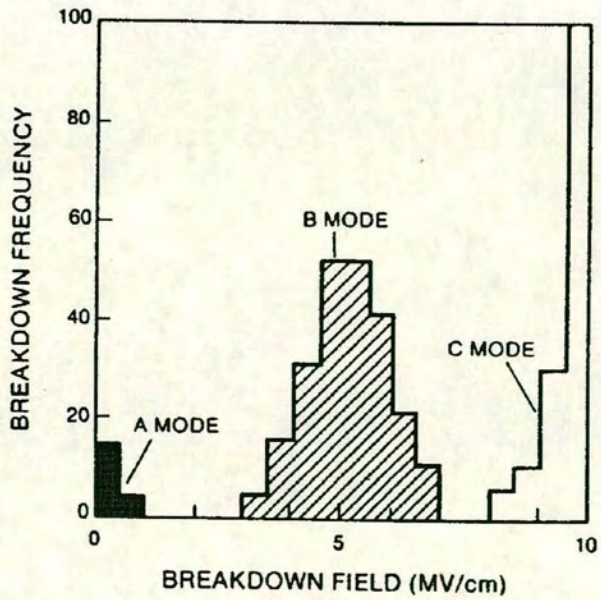


Figure 4.4 Histogram of dielectric breakdown electric field for a number of thin oxide capacitors showing the three distributions referred to as Mode A, B and C fails [41].

4.2 Oxide Breakdown Phenomenon

Thin oxides breakdown such that a path of conduction exists through the gate oxide insulator when they are stressed under high electric fields. Not all thin oxide capacitors breakdown at the same electric field. Figure 4.4 shows the typical distributions of capacitor fails referred to as Mode A, B and C.

Mode A fails are associated with yield problems where the thin oxide usually has a conductive path through prior to voltage stress as a result of microcontamination during the oxide growth or polysilicon deposition, or from Electrostatic Discharge (ESD) from equipment or wafer handling.

Mode B fails are attributed to oxides breakdown at intermediate electric fields (2-6 MV/cm). These oxides present the most concern as they show little if any degredation at operating voltages but will degrade over longer periods of time at operating voltages. These fails are also referred to as time dependant dielectric breakdown (TDDB) fails or extrinsic breakdown fails.

Finally Mode C fails occur in the 8-12 MV/cm range and are associated with the intrinsic breakdown of the thin oxide. Intrinsic breakdown implies that the oxides will perform reliably beyond the operating lifetime of the chip. It is obviously desirable to have a large population of capacitors sampled for stressing be Mode C fails.

The exact nature of oxide breakdown is still not certain. There are have been many models proposed to describe this phenomenon such as the electron-trapping model, the resonant tunnelling model and the electron-lattice damage model [42,43,44]. In this section the most common models are described.

4.2.1 Hole Generation and Trapping Model

The hole generation and trapping model is the most widely accepted model used to describe intrinsic oxide breakdown [45]. The basis of the model is that at high electric fields, electrons are injected into the oxide conduction band through FN tunnelling where they are subsequently accelerated towards the anode. These high kinetic energy electrons cause the formation of electron-hole pairs in the oxide. A small fraction of the generated holes accumulate in areas of the oxide where imperfections exist and the local electric field is increased. This increased electric field in localized areas causes more current to tunnel through the oxide and the positive feedback continues until the point of breakdown. Oxide breakdown is considered to be a two stage process. Firstly the oxide is damaged by the hole injection into the oxide over different periods of time depending on the electric field. The second stage is the short run away process caused by the positive feedback. The time to failure is determined by the first stage of the process.

The trapping of holes is considered to be the precursor to oxide damage. Holes have a much lower mobility in the oxide than electrons and are therefore more likely to be trapped than electrons. Hole traps are thought to occur from oxygen vacancies in the bulk oxide or dangling bonds at the oxide surface. Although electrons can be trapped throughout the oxide, this model assumes that electrons do not contribute to oxide damage. Figure 4.5 shows the energy-band diagram for the case of both electron and hole trapping. It is evident from this diagram that hole trapping reduces the FN tunnelling barrier more than electron trapping.

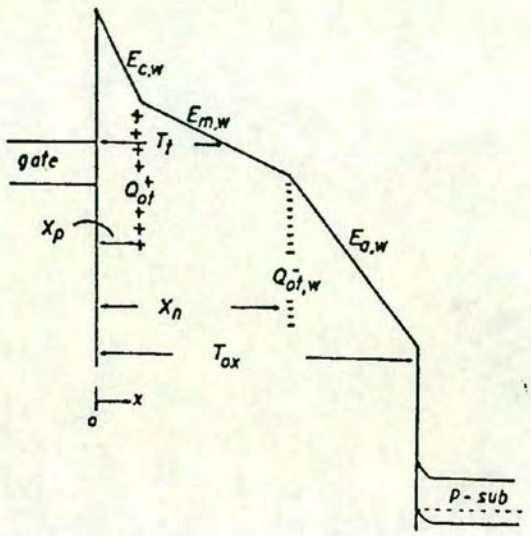


Figure 4.5 Energy-band diagram showing effect on the FN tunnelling barrier from trapped hole charge Q_{ot}^+ and trapped electron charge Q_{ot}^- [46].

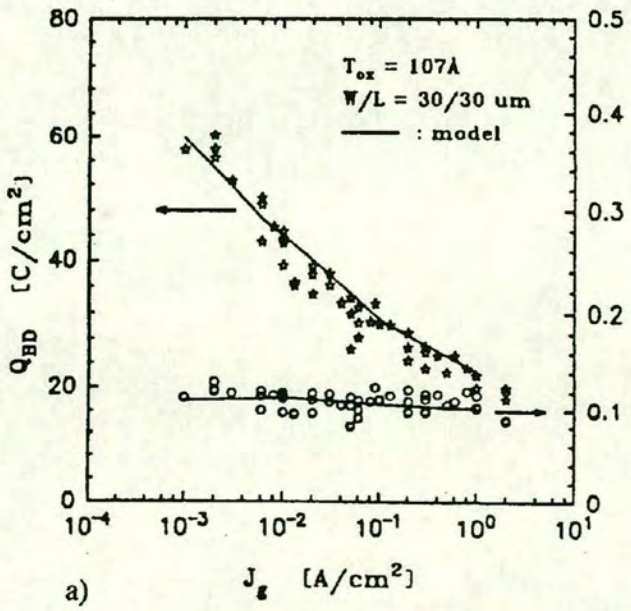


Figure 4.6 Charge to breakdown from electron trapping Q_{BD} and charge to breakdown from hole trapping Q_P verses injected gate current for a 107Å oxide [47].

In an experiment to verify that holes are responsible for oxide damage, a NMOSFET with 107Å oxide thickness was stressed with a positive bias applied to the gate. Figure 4.6 shows the charge to breakdown from the electron gate current Q_{BD} and the charge to breakdown from the substrate hole current Q_P. For all of the of gate current densities studied, the hole charge to breakdown is lower than the electron charge to breakdown.

One problem with this model is that the charge to breakdown decreases with increased temperature. Experimental data shows that the charge to breakdown increases with increased temperature. One possible explanation for this discrepancy is that hole trapping causes localized bandgap narrowing in the oxide and at low temperatures barrier lowering by the electric field is required for holes to be trapped. This also explains why Q_P increases at low gate current densities for low temperatures.

4.2.2 Low Voltage Intrinsic Breakdown Model

Recently a model has been proposed [48] to describe the low voltage breakdown of thin (<150Å) oxides. In this model the energetic electrons tunnel through the oxide and upon reaching the anode, they transfer their energy to deep valence band electrons. The deep valence band electrons are promoted to the conduction band which produces a hot hole in the valence band. These hot holes increase the local electric field by the previously mentioned hole induced trap generation model until breakdown occurs. Figure 4.7 shows the energy band diagrams for this anode hole injection process.

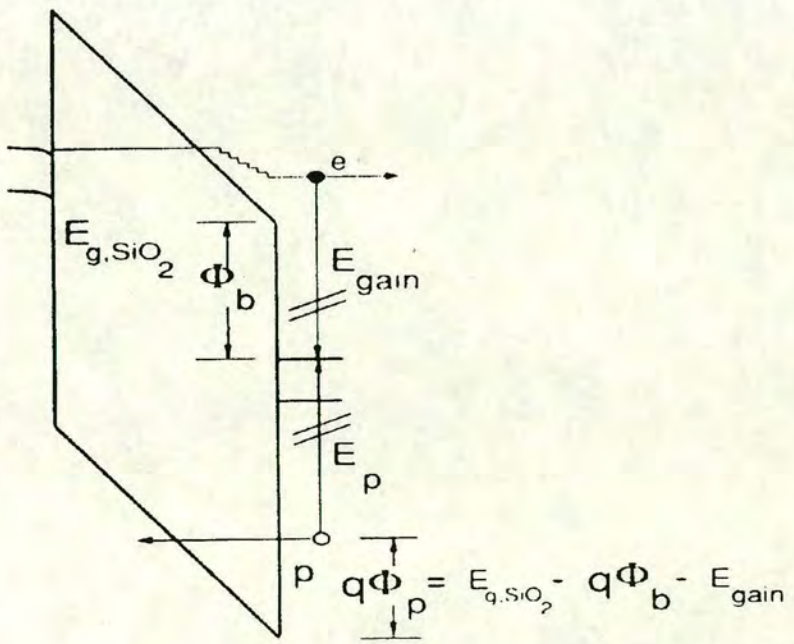


Figure 4.7 Energy band diagram showing the anode hole injection process [48].

4.2.3 Empirical Model for Intrinsic Oxide Breakdown

This empirical model [49] was developed due to the ongoing debate concerning the mechanism for oxide breakdown. This model is a statistically based model generated from experimental evidence. The basis for the model is that at relatively small FN tunnelling currents, the Q_{BD} of intrinsic oxides is constant at ~10C/cm^2 and Q_P is also constant at ~0.1C/cm^2. The time to breakdown t_{DB} can be modelled as

$$Q_P \sim J_g \alpha t_{DB} \qquad 4.2$$

where α is the hole-generation coefficient and J_g is the gate current predicted using Equation 4.1. The hole-generation coefficient is given by

$$\alpha \sim \exp\left(-\frac{H}{\varepsilon_{ox}}\right) \qquad 4.3$$

Using Equations 4.1, 4.2. and 4.3, the time required to accumulate the charge Q_P to cause oxide breakdown t_{BD} is

$$t_{BD} \propto \exp\left(\frac{G}{\varepsilon_{ox}}\right) \qquad 4.4$$

where

$$G = H + B \cong 350 MV cm^{-1} \qquad 4.5$$

The addition of temperature dependence to Equation 4.4 then becomes

$$t_{BD} = \tau_o(T) \exp\left[\frac{G(T)}{\varepsilon_{ox}}\right] \qquad 4.6$$

where τ_o is the pre-exponential factor. At 300K, Equation 4.6 reduces to

$$t_{BD} = 10^{-11} \exp\left(\frac{350 \cdot t_{ox}}{V_{ox}}\right) \qquad 4.7$$

Figure 4.8 shows measured data and the model fit for oxides of different thickness with an n^+polysilicon gate on a p-type substrate. Although the relationship between $\log(t_{BD})$ and $1/\varepsilon_{ox}$ is linear, their is a maximum electric field where this relationship is no longer valid. Accelerated tests should therefore be carried out at a lower than maximum electric field when using this model.

Equation 4.6 is the basis for extrapolating oxide lifetime (t_{BD1}) at higher than normal stress voltage (V_{ox1}) to lifetime (t_{BD2}) at normal operating voltage (V_{ox2}).

This equation is written

$$t_{BD2} = t_{DB1}\left(\frac{\tau_o}{t_{BD1}}\right)^{\left(1-\frac{V_{ox1}}{V_{ox2}}\right)} \qquad 4.8$$

A point of note is that although V_{ox} is the voltage on the gate electrode, the actual voltage over the oxide is significantly less. This is due to the polysilicon depletion effect which results from band bending in the polysilicon. Since the band bending is a function of polysilicon doping and work function differences exist between n^+ and p^+ polysilicon gates the amount of voltage loss will vary. This model is only valid for small area capacitors which are considered to be defect free.

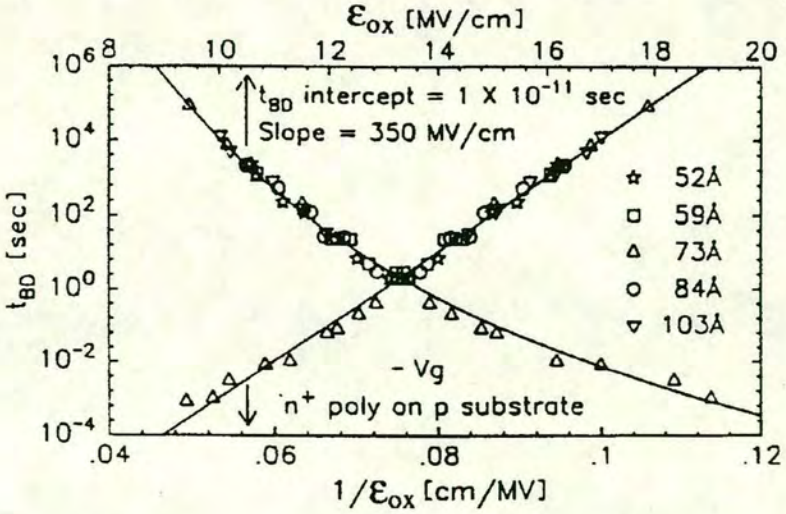


Figure 4.8 $\text{Log}(t_{BD})$ verses $1/\varepsilon_{ox}$ for different gate oxide thicknesses. A good agreement is found between the measured data and the model [49].

4.2.4 Physical Models for Mode B Fails

As mentioned previously, Mode B oxide fails are of most concern because they will cause transistors to fail at operating voltages after long periods of time. These Mode B type oxide fails are believed to be caused by defects in the oxide which have not caused catastrophic conductive paths through the oxide. Several physical mechanisms have been found to cause these 'walking wounded' oxides.

Contamination by metallic elements such as Fe, Al and Ca, on the silicon prior to oxide growth or during oxide growth can degrade the oxide properties [50]. These metallic contaminants can originate from cleaning chemicals or processing equipment. Mobile sodium ions at high enough concentrations can cause localized high electric fields which can degrade the oxide. The tungsten filament in furnace tubes contains levels of sodium which can migrate through the quartzware to the silicon wafers. Periodic cleaning of quartz can prevent this problem.

Microscopic defects generated in the silicon crystal lattice can propagate to the silicon surface and cause a weak oxide to be grown. The source of the defects in the silicon has been attributed to the presence of oxygen during the silicon crystal growth. The out diffusion of oxygen from the silicon surface by denuding can alleviate this problem [51].

Surface microroughness can cause weak oxides to be grown [52]. As was mentioned in Chapter 2, the specific cleaning solutions used to remove the native oxide, metallic and organic contaminants can cause varying amounts of micropitting on the silicon surface. The silicon surface microroughness has been shown to be translated up to the surface of as-grown thin oxides. The technique of Atomic Force Microscopy (AFM) [53] can measure the degree of roughness. AFM involves the scanning of a stylus tip over the silicon surface. The stylus tip is held with a constant force over the wafer by a piezoelectric positioner. The voltage applied to the positioner to maintain the constant force is measured as the stylus moves over the wafer and a three dimensional representation of the silicon surface is obtained.

The oxide breakdown can be lowered as a result of localized thinning of the oxide. The thinning can be as a result of stress effects or the Kooi effect. Stress effects are common where the corner of a silicon trench is oxidized [54]. The Kooi effect is the formation of silicon nitride layer on the silicon surface during LOCOS [55]. The common way to solve the Kooi nitride effect is to strip the grown oxide and regrow a higher quality one. This sacrificial oxide is usually implanted through during threshold adjust implants and would have ion implant damage.

As gate oxides are made thinner for performance reasons, it becomes difficult to prevent

significant thinning from the polysilicon gate etch. Although with modern equipment it is still possible to get a good etch selectivity of polysilicon to oxide, damage can still occur from the scattering of ions during the over-etch from the etched polysilicon sidewall [56]. These preferentially etch areas extend through the oxide and into the silicon substrate at the polysilicon gate edge and are called microtrenches. Optimization of the polysilicon etch can reduce this effect. Thin oxide can also be damaged from antenna effects during plasma processing [57,58]. This problem will be covered in more depth in section 4.4.

Since it is more likely to find a defect in a large capacitor, small capacitors are used to measure the intrinsic oxide breakdown. In order to establish the density of defects and to monitor process improvements, large capacitors are required.

4.2.5 Qualitative Models for Mode B Fails

The oxide thinning model [49] is an extension of the model described in Section 4.2.3 to account for extrinsic oxide failures. Defects in the oxide which cause premature oxide failure are modelled by assuming that the oxide is of intrinsic quality but effectively thinner. The thinnest oxide limits the oxide breakdown voltage. Figure 4.9 shows the oxide thinning with the inclusion of a defect. The defect is assumed to be conductive so that there is no voltage across it. Equation 4.6 is modified to account for the effective thickness of the oxide t_{oxeff} to give

$$t_{BD} = \tau_o(T)\exp\left[G(T)\frac{(t_{ox} - \Delta t_{ox})}{V_{ox}}\right] = \tau_o(T)\exp\left[G(T)\frac{t_{oxeff}}{V_{ox}}\right] \qquad 4.9$$

where Δt_{ox} is the change in effective oxide thickness from the presence of a defect. A Poisson distribution can be used to describe the probability distribution of t_{BD}' resulting from a uniform random distribution of defects on the wafer with oxide thinning Δt_{ox}. Equation 4.10 below gives the probability distribution of t_{BD}' where A is the oxide area and $D[\Delta t_{ox}]$ is the cumulative defect density function.

$$P(t_{BD}' < t_{BD}) = 1 - \exp(-AD[\Delta t_{ox}]) \qquad 4.10$$

A gamma function can be used to describe distribution of t_{BD}' resulting from the nonuniform distribution of defects on the wafer. S is the degree of clustering and for the case of no

clustering, S=0 and Equation 4.11 reduces to Equation 4.10.

$$P(t_{BD}' < t_{BD}) = 1 - \frac{1}{(1 + AD[\Delta t_{ox}]S)^{1/S}} \qquad 4.11$$

Figure 4.10 shows the cumulative defect density function $D[\Delta t_{ox}]$, obtained from the Gamma function for various oxide thicknesses.

The increased trapping model is another extension of the model described in Section 4.2.3. This model accounts for Mode B failures by altering the pre-exponential factor and has been postulated to occur as the result of an increased concentration of trapping centres in the oxide. Equation 4.6 then becomes

$$t_{BD} = \tau_{oeff}(T)\exp\left[G(T)\frac{t_{ox}}{V_{ox}}\right] \qquad 4.12$$

τ_{oeff} is the modified pre-exponential factor.

A comparison between these two models has shown that the effective oxide thinning model fits the experimental results better [59].

4.3 Test Methods to Establish Thin Oxide Reliability

Oxide breakdown can be measured using various methods. To determine the intrinsic oxide breakdown distribution a large sample of small area capacitors are required as the level of defectivity should be low. On the other hand to establish the defect density resulting in extrinsic failures, a large sample of capacitors are required with a wide range of capacitor areas and layout styles. In order to get results in a relatively short time (seconds), time zero dielectric breakdown (TZDB) tests are usually performed. There are however studies which have shown that the mechanism responsible for oxide breakdown during highly accelerated wear-out tests is different from that at operating voltages. Time dependent dielectric breakdown (TDDB) is used to determine the oxide reliability at a lower acceleration factor where the mechanism responsible for breakdown is considered to be the same as that at operating voltages. TDDB tests however usually take several months to complete and hence result in a long time for oxide quality improvement programs to complete.

In practice, a combination of TZDB and TDDB is used. The TZDB is usually used to provide an ongoing monitor of oxide quality and as a means of monitoring process modifications for improved oxide integrity. TDDB is then used to determine the lifetime of an oxide at operating voltages once the process has been optimised.

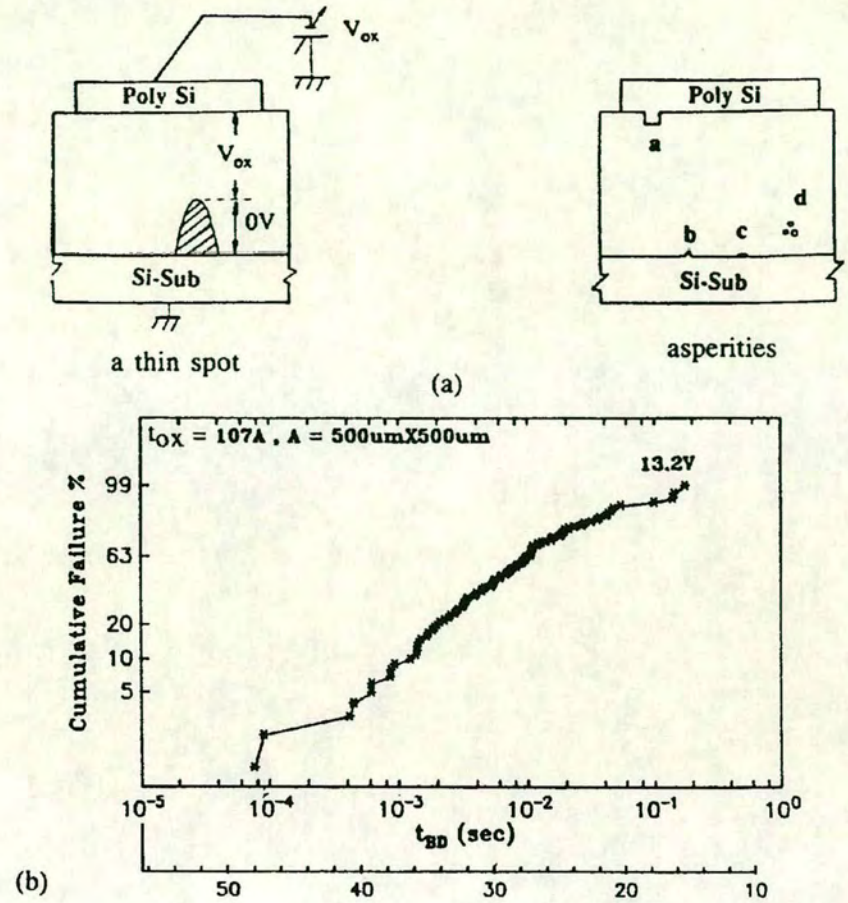


Figure 4.9 Defects in the oxide are modelled as effectively thinning the oxide [49].

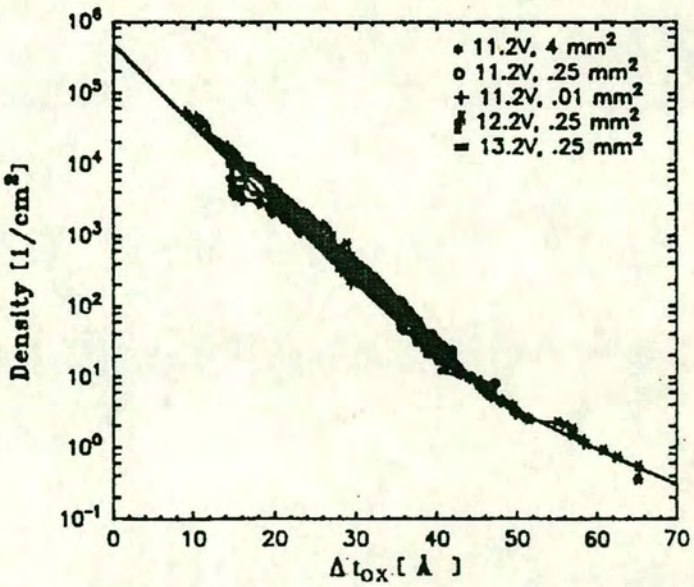


Figure 4.10 Defect density as a function of Δt_{ox} using the Gamma distribution [49].

4.3.1 Time Zero Dielectric Breakdown

Time Zero dielectric breakdown tests [60] can be performed using either a current or a voltage ramp. A linear or logarithmic ramp of the current is usually used to determine major problems with oxide integrity. This can be seen by considering Equation 4.1 where for low values of ε_{ox}, there is a high resolution of J. Oxide breakdown is considered to be the point at which the measured voltage over the capacitor suddenly drops. The charge to breakdown Q_{BD}, can be easily calculated by this method by integrating the stressed current over time [61].

For oxides which are expected to be close to intrinsic, a voltage ramp method is recommended to obtain good resolution of the FN tunnelling current. Figure 4.11 shows the typical staircase ramp of the stress voltage and the corresponding measured leakage current. Oxide breakdown voltage is dependent on the rate of voltage ramp. For the purposes of DEC microprocessor designs, 1μA is considered to be a significant enough leakage current to cause circuit problems [59]. While it is common practice to bias the p-type substrate capacitor in accumulation mode, there is not a general acceptance as to how the n-type substrate capacitor should be biased. This is due to the voltage drop over the depletion layer when the n-type substrate is biased in inversion. This voltage drop over the unwanted series capacitor tends to superficially increase the intrinsic breakdown field of the oxide [62]. If the n-type substrate capacitor is biased into accumulation however, then the poorer quality polysilicon-silicon dioxide interface is where electrons are injected into. This results in a lower intrinsic breakdown voltage than would be expected and the added variability lot to lot from the polysilicon deposition morphology. In this work the n-type substrate capacitors are biased into inversion to increase the oxide breakdown sensitivity to the quality of the silicon-silicon oxide interface.

4.3.2 Time Dependent Dielectric Breakdown

The application of a fixed voltage over time [63] is the most common method of time dependent dielectric breakdown (TDDB). The applied stress field is held constant over a long period of time and the time to breakdown is when the oxide leakage current meets a certain value. Although a high stress field results in a shorter time to fail, it is considered more accurate to stress the oxide as close to operating voltages as possible. Since oxide breakdown is a function of stress temperature, capacitors are often stressed around 125°C as a means of maintaning the same acceleration factor at a lower voltage. TDDB is performed at a packaged level compared to TZDB which is performed at the wafer level.

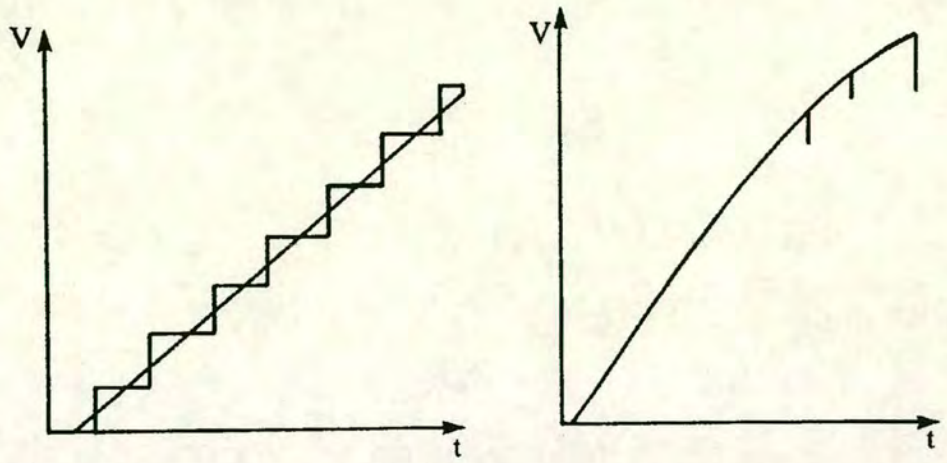


Figure 4.11 (a) Typical staircase voltage ramp applied to the capacitor. After each incremental voltage increase, the stress voltage is held and the capacitor leakage is sensed to determine the point of breakdown. (b) Resulting FN tunnelling current and point of breakdown determined at 1μA [59].

4.3.3 Projecting Accelerated Wear-out Tests to Operating Voltages

As it is not practical to measure the lifetime of oxides at operating voltages, it is necessary to use an accelerated test method. A correlation has been found [64] between TZDB and TDDB tests which can be used to extrapolate both tests to operating conditions. For a constant voltage stress test at voltage V_{ox} and temperature T_{tbd} resulting in a time to fail t_{BD}, and a ramped voltage test at room temperature T_{vbd} with ramp rate R and failure voltage V_{BD}, Equation 4.13 applies.,

$$t_{BD} \cong \tau_o(T_{tbd}) \left| \frac{V_{BD}^2}{RG(T_{tbd})\tau_o(T_{vbd})t_{ox}} \right|^{G(T_{tbd})V_{BD}/G(T_{vbd})V_{ox}} \qquad 4.13$$

This equation predicts that an oxide with a lifetime of 10 years at 5.5V and 125°C will fail in a 1V/sec voltage ramp test at V_{BD}~11.5V and a that a lifetime of 100 years at 125°C will correspond to a V_{BD}~12V at 1V/sec.

4.3.4 Burn-in Pre-screening

Inorder to eliminate mode B failure types, which would pass D.C. testing of a transistor and could cause failure of a chip during normal operation, burn-in pre-screening is applied. Burn-

in involves a relatively short (~1000sec) high temperature high voltage stress of packaged circuits prior to shipment to the customer. The burn-in conditions are selected based on the distribution of mode B fails found by stressing a small population of transistors to destructive breakdown [64]. The burn-in will also effect the intrinsic oxides, the operating lifetime t_{op} will then be

$$t_{op} = \tau_o(T_{op})\exp[G(T_{op})t_{oxeff}/V_{op}L_{BI}] \qquad 4.14$$

where T_{op} is the operating temperature and V_{op} is the operating voltage. L_{BI} is the fraction of lifetime remaining after burn-in, for t_{BI} seconds at burn-in voltage V_{BI} and temperature T_{BI}.

$$L_{BI} = [1 - t_{BI}\exp -G(T_{BI})(t_{oxeff}/V_{BI})] \qquad 4.15$$

4.3.5 Scaling Failure Probability Rate by Area

The negative binomial model can be used to scale failure rates between different test structure areas or between test structure area and circuit area. The scaled failure probability P of the circuit is given by

$$P = 1 - [A_c/A_t(Y_t^{-1/\alpha} - 1) + 1]^{-\alpha} \qquad 4.16$$

where A_c and A_t are the circuit and test structure areas (or periphery length) respectively, Y_t is the yield of the test structure and α is the clustering coefficient

4.4 Gate Oxide Damage from Plasma Processing

As was mentioned in Section 4.2.4, thin gate oxides can be damaged from the use of plasma sources in the fabrication of integrated circuits. The damage results from non-uniformities in the plasma across the wafer surface which can result in significant charge build-up [65]. Although the net current flow over the wafer as a whole is balanced, the local electron and ion currents can vary. In the case when the surface of the wafer is a conductor, such as doped polysilicon or metal, the surface currents flow to balance the local non-uniform currents. If the conducting material is situated upon an oxide, the area of the conducting material will act like an antenna and the collected charge will be situated over the oxide to the grounded substrate. When the collected charge becomes significant compared to the Q_{BD} of the oxide, the oxide will be damaged through FN tunnelling. This type of charging damage can cause intrinsically good oxides to become weak or even shorted. Both positive and negative

charging can occur with voltages over 50V possible on thick oxides in very non-uniform plasma's. It is common to have the conducting antenna over both thick field oxide and thin gate oxide regions. Since the charge sharing will result in more charge over the thinner oxide, the funnelling of this charge is referred to as the Antenna Effect [66]. The Antenna Area Ratio (AAR) is defined as the surface area of the conducting material over the thick oxide divided by the surface area over the thin oxide. In the case where the charge is collected only on the sidewalls of the conductor, the AAR is then the ratio of the conductor periphery area to the thin oxide area.

4.4.1 Models for Thin Oxide Charging from Plasma Processing

Fang and McVittie have proposed a model [65] to explain oxide charging from RIE etching of the doped polysilicon gate material. As the bulk of the doped polysilicon is etched in the areas not covered by photoresist, the nonuniform currents balance out as the whole wafer is still connected in polysilicon and there is therefore no damage to the oxide. When the island of the gate electrode start to appear close to end-point, the etched polysilicon thickness becomes so thin that the path of least resistance for the collected charge now becomes through the oxide and serious damage can occur. This charge build up continues through to the overetch step where the charging is limited to the area of the polysilicon sidewall. These steps in the etching of the polysilicon electrode are shown in Figure 4.12. Damage to the oxide has been found to increase as the ratio of gate perimeter to area increases. The type of damage described in Fang and McVittie's model also occurs during the gate spacer etch and during the etching of aluminium conductors.

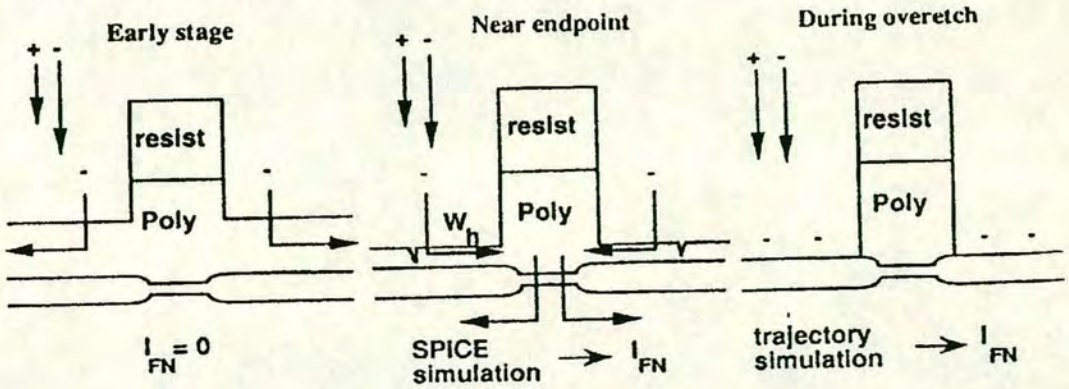


Figure 4.12 Charge damage to the gate oxide during polysilicon etching. Initially the surface currents through the polysilicon prevent charging. Close to endpoint, the thin polysilicon resistance becomes high and the resulting charge build up is discharged through the oxide. During overetch, only the polysilicon sidewalls are charged and the additional amount of damage to the oxide is reduced [65].

Plasma charging can also occur during the stripping of photoresist from patterned conductors. In this case the charging occurs at the conductor edges during the photoresist strip and over the conductor surface as a whole during endpoint and overetch. During the overetch of contacts through an oxide down to a conductor significant plasma charging can also occur. This effect is especially bad on large area bond pads where the whole pad is open to the etch.

Other process steps which can cause oxide damage in nonuniform plasma are pre-metal sputter cleaning, Plasma Enhanced-CVD oxide deposition and ion implantation.

4.4.2 Methods to Monitor Plasma Induced Oxide Degredation

Plasma charging effects which affect MOS transistors can be amplified on test structures which have a large area of thin oxide or structures which have a high AR. Large area capacitors are more likely to contain a plasma induced weak oxide due to the same defect density distribution found from particles in the oxide. It is difficult however to distinguish plasma damaged oxides from particle defectivity. Antenna structures of varying AR can be used to monitor process changes and quantify the extent of the plasma charging [67]. The antenna can be an area or a periphery intensive structure depending on the process step being investigated as a source of oxide damage and it is connected to the gate of a MOS transistor.

The generation of interface states by the FN tunnelling current associated with plasma charging will induce a threshold voltage shift in the MOS device. Measurements of the difference in threshold voltages between a MOS transistor with and without a certain antenna will give an indication of the degree of damage suffered by the thin oxide. The C-V technique can also be applied to measure the increase in interface traps from plasma etch steps compared to a wet etch control.

Other techniques used to measure the damage to the oxides from plasma charging include voltage ramp dielectric breakdown and charge-to-breakdown. The cumulative oxide breakdown of a sample of transistors with varying antenna ratios are usually compared with a sample of transistors with the same antenna structures but with a protective diode introduced at various steps throughout the process. This comparative study would enable the source of the most damage to be narrowed down to a particular process step. The damaged caused by plasma processes has been shown to have a cumulative effect throughout the process so there is more to gain by identifying and fixing the processes which cause the most damage.

4.4.3 Methods to Reduce the Plasma Charging Effect

An obvious way to reduce the effects of plasma charging of thin oxides is to optimize the plasma conditions to obtain uniform electron and ion currents. This can be made easier by

choosing a plasma system with a low excitation frequency and in the case of magnetically enhanced RIE systems, optimizing the magnet configuration [68]. The use of downstream resist strippers can reduce the damage caused from ashing. In this case, the plasma is generated in a remote location from the wafer and only the neutral charged chemically active species are exposed to the wafer.

Another approach to reduce the plasma damage is to cover the conducting antenna with a protective dielectric and pattern the electrode by etching both the dielectric and conductor at the same time [69]. In the case of oxide damage resulting from the contact etching over a bond pad area, a simple solution is to etch lots of small contacts to reduce the exposed conductor area.

The most common way of protecting thin oxides to plasma damage is to connect a diode in parallel with the MOS device so that the charge flows through the diode to the substrate [70]. This type of protection however only protects the NMOS device from negative charge build-up with a n^+/p diode and the PMOS from positive charge build-up with a p^+/n diode. Since studies have shown that the protecting diode can be damaged and to limit the damage from the other type of charge build-up, antenna layout design rules are usually added.

CHAPTER 5

MOSFET Hot Carrier Effects and Related Device Engineering

The topic of hot carrier effects are discussed in this chapter in order to describe the various methods of measuring and improving the robustness of MOS devices and to assess the potential improvements in hot carrier lifetime with the molecular nitrogen implanted silicon technique.

5.1 The Hot Carrier Effect in N-MOSFET Devices

5.1.1 Mechanism for Hot Carrier Damage.

The shrinkage of MOSFET dimensions into the submicron regime at constant supply voltage increases the hot carrier effect. This phenomenon arises from the increased lateral electric field in the channel as the channel length is reduced. The inversion layer charges are accelerated by the lateral electric field to a point at which they damage the device. Figure 5.1 shows the hot carrier generation and resulting current components associated with channel hot electron (CHE) effects.

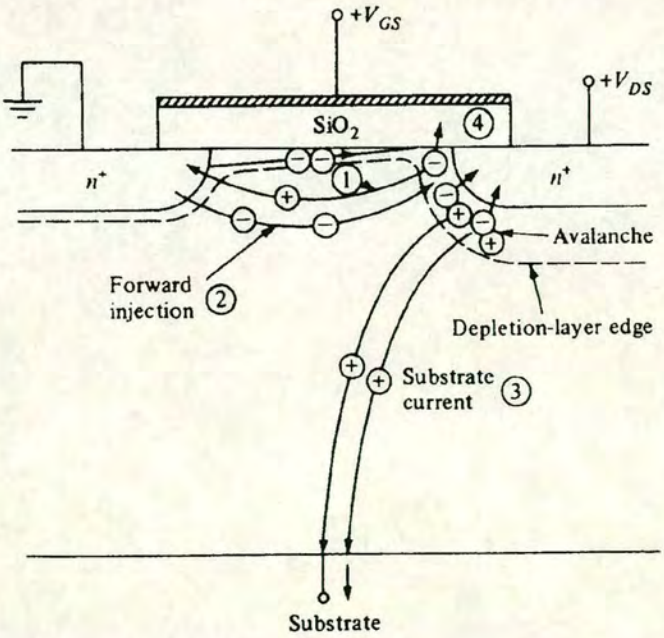


Figure 5.1 Generation of hot carriers and resulting current components in a N-MOSFET. 1. Holes reaching the source. 2. Electron injection from the source. 3. Substrate hole current. 4. Electron injection into the oxide [71].

The most damaging current component resulting from the hot carrier generation is the injection of charges into the Si/SiO_2 interface or into the oxide. Important device parameters

vary with time from hot carrier effects as discussed in Section 5.2. There are methods used to reduce the damage caused by hot carrier effects in submicron MOSFET's. An obvious solution would be to reduce the operating voltage and hence scale the electric field in the channel, this method however reduces the operating speed of the device and would be impractical for sub-half micron channel lengths. Since the degree of damage is associated with both the location and magnitude of the maximum electric field, another approach to reduce the damage is to re-engineer the source/drain profiles. This topic will be covered in depth in Section 5.3. The gate oxide and silicon dioxide-silicon interface can be hardened to hot carrier effects, this topic will be addressed in Section 5.4.

5.1.2 Models for the Maximum Electric Field in N-MOSFET's.

The maximum lateral electric field ε_{ymax} is located near the drain in non-graded drain MOSFET's. The hot carrier effects only become significant when ε_{ymax} becomes greater than $4x10^4$ V/cm which occurs when the carriers reach the velocity saturation regime [72]. An early two-dimensional numerical solution to Poisson's equation results in the electric field verses the lateral position in the channel as shown in Figure 5.2 for two different channel lengths [73].

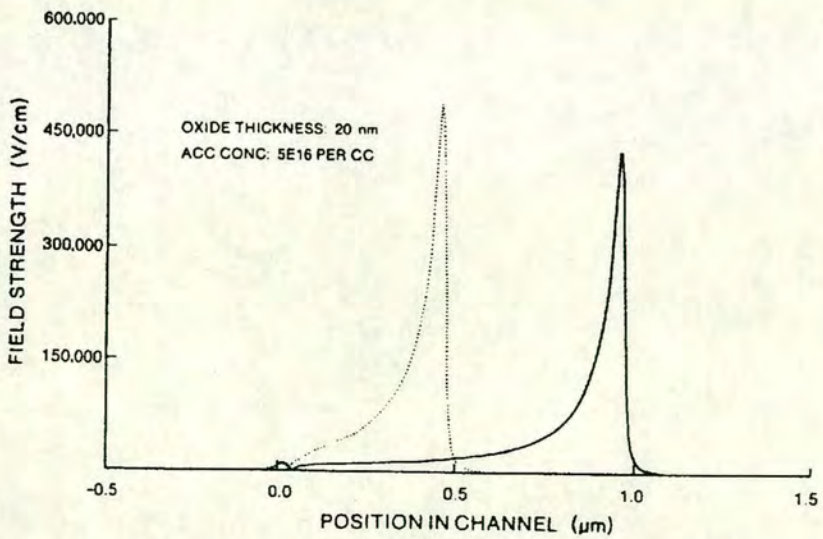


Figure 5.2 ε_y verses channel length for a 0.5μm (dotted line) and a 1.0μm (solid line) channel length N-MOSFET with t_{ox}=20nm [73]

In the resulting model, the maximum lateral electric field is given by Equation 5.1.

$$\varepsilon_{ymax} = \frac{(V_{DS} - V_{DSsat})}{l} \qquad 5.1$$

for the case when

$$(V_{DS} - V_{DSsat})/l \gg \varepsilon_{sat} \qquad 5.2$$

The source/drain junction depth r_j, gate oxide thickness t_{ox} and channel length L, were numerically analysed such that the value l is given by

$$l = 0.22 t_{ox}^{1/3} r_j^{1/3} \qquad t_{ox} \geq 15nm \qquad 5.3$$

or

$$l = 1.7 \times 10^{-2} t_{ox}^{1/8} r_j^{1/3} L^{1/5} \qquad t_{ox} < 15nm, L < 0.5\mu m \qquad 5.4$$

The above equations indicate that for thick gate oxides, the term l is independent of L until L is close to the value of l for a given technology. The term V_{DSsat} which is the saturation drain voltage, is expressed as [74]

$$V_{DSsat} = \frac{\varepsilon_{sat} L [V_{GS} - V_T]}{(V_{GS} - V_T) + \varepsilon_{sat} L} \qquad 5.5$$

The effect of oxide thickness on the lateral electric field is shown in Figure 5.3 below.

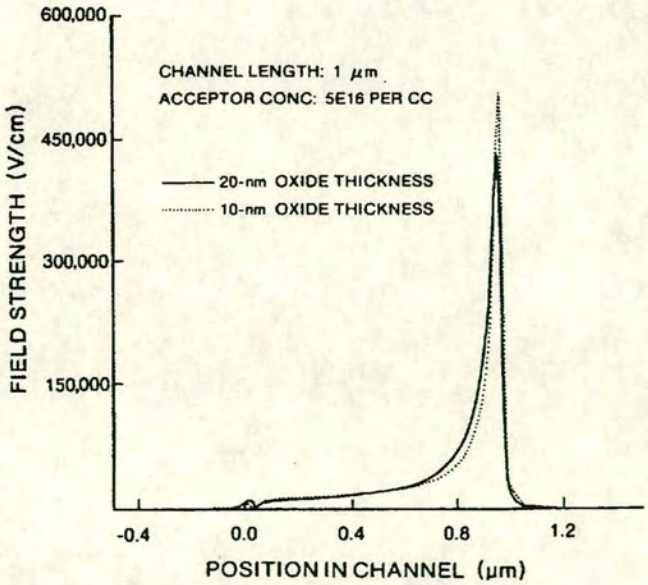


Figure 5.3 ε_y verses channel length for a 1.0μm N-MOSFET with gate oxide thickness 10nm and 20nm. [73]

5.1.3 Substrate Current Monitor of CHE.

The substrate current which is generated in a MOSFET (See Figure 5.1) from CHE can cause damage to the device by; forward biasing the Source/Substrate diode causing the device to go into snapback, inducing latchup, shifting the substrate bias condition or causing secondary impact ionization far from the drain junction thereby causing leakage currents in Dynamic Random Access Memories (DRAM) circuits [75].

Carriers in a silicon device are termed 'hot' if they posses energies in excess of 1.5eV, which is the point of thermal equilibrium in the silicon lattice. In a N-channel MOSFET, hot carriers are generated from impact ionization in the channel which results from the collision of channel electrons with the lattice. The resulting electron-hole pair splits up such that the electron is accelerated towards the drain or into the gate oxide if it possess a energy greater than the Si-SiO_2 barrier height of 3.2eV. The generated holes are attracted to the substrate and form the substrate current. Since there is a hole created for every hot electron, the substrate current can be used as a monitor to correlate device degradation with predicted device lifetime. The measured substrate current I_{SUB} verses applied gate voltage V_{GS}, for different V_{DS} values is shown in Figure 5.4. There can be seen to be a point at which the substrate current is a maximum which corresponds to the case of $V_{GS} \approx 0.4 V_{DS}$

The substrate current in Figure 5.4 increases at low gate voltages as there is an increase in the number of channel electrons which can participate in impact ionization. As the gate voltage increases towards the device saturation condition the substrate current decreases due to a reduction in the lateral electric field while the drain current continues to increase. There is however an effect called Drain Avalanche Hot Carriers (DAHC) which occurs in devices with thin oxides at high drain voltages where hot electrons have a greater likelihood of reaching the gate as the potential difference between the gate and drain is lowered [76]. DAHC also reduces as the gate voltage approaches the drain voltage due to the reduced incidences of impact ionization from a lower lateral electric field. The NMOS gate current is therefore due to hole current apart from the DAHC electron injection.

Figure 5.5 shows the dependence of the measured maximum substrate current of the channel length. Thinning the gate oxide and shallowing the source/drain junctions will also increase the substrate current according to Equation 5.4.

Figure 5.6 shows the straight line dependence of the logarithm of substrate current with the reciprocal of drain voltage which was found to be independent of gate voltage or channel length. As a result of this finding, an analytical model [77] was derived to calculate the

substrate current as follows

$$I_{SUB} = 1.2(V_{DS} - V_{DSSat})I_D \exp\left(\frac{-3.7 \times 10^5 t_{ox}^{1/3} r_j^{1/3}}{V_{DS} - V_{DSSat}}\right) \qquad 5.6$$

The above equation is valid only for a uniformly doped channel transistor.

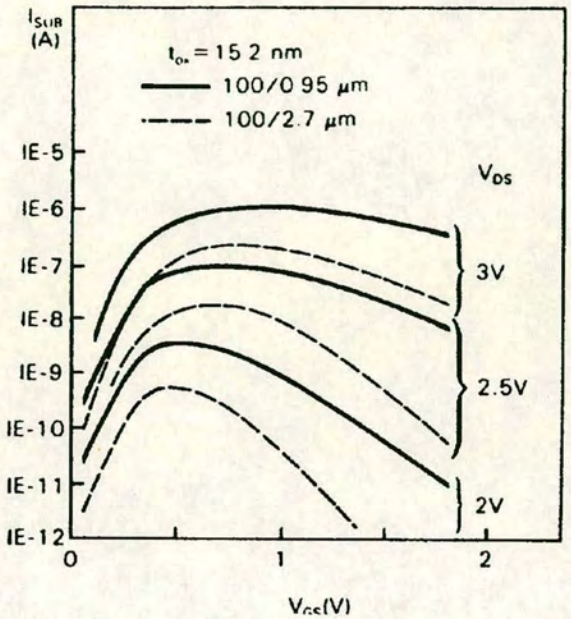


Figure 5.4 Measured substrate current verses gate voltage for different drain voltage conditions [77].

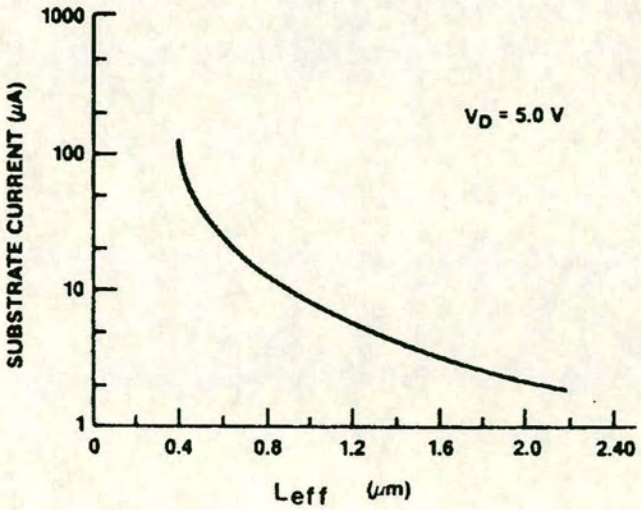


Figure 5.5 Maximum substrate current as a function of channel length for a 5V drain voltage and a t_{ox}=250Å [78].

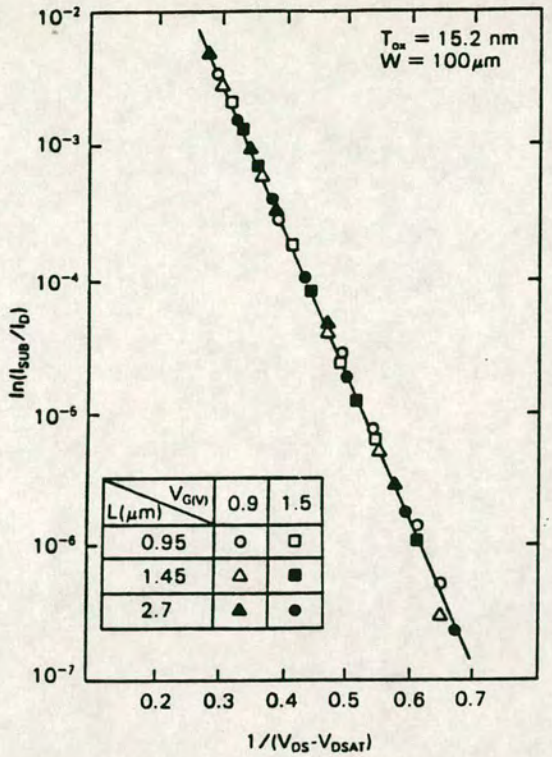


Figure 5.6 Logarithm of substrate current verses reciprocal of drain voltage as a function of gate voltage and channel length [77].

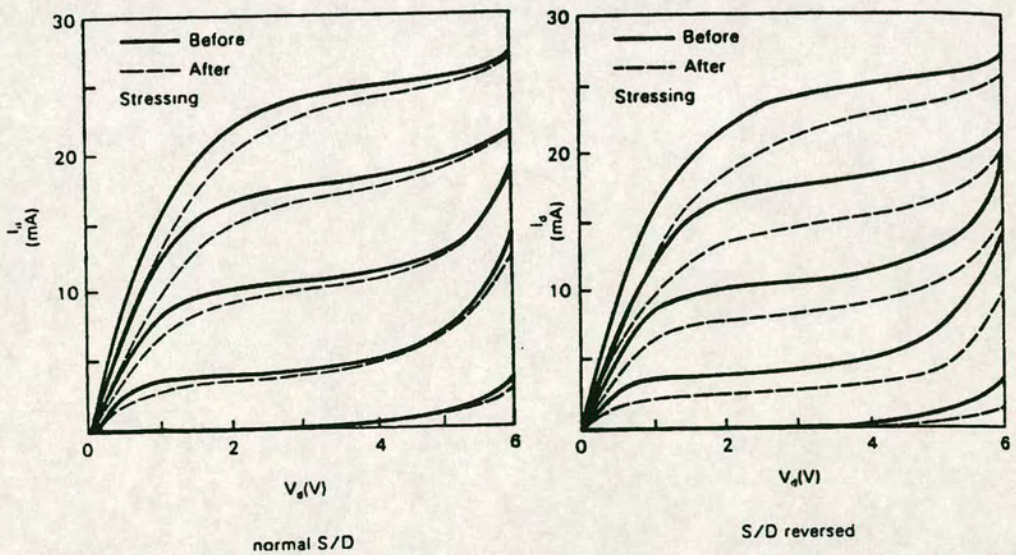


Figure 5.7 Typical device characteristics before and after hot carrier stress [79].

5.1.4 Models for Hot Carrier Degredation in N-MOSFET's

There are several models that have been proposed to describe how hot carriers degrade N-MOSFET's but there is still no consensus as to which model is correct. An examination of an N-MOSFET operating characteristics before and after hot electron stress as in Figure 5.7, indicates that a build up of negative charges in the gate oxide or at the Si-SiO_2 is responsible for increasing the threshold voltage, decreasing the transconductance and reducing the drive current [80]. The result of reversing the source/drain terminals to examine the operating characteristics after hot carrier stress shows a more pronounced effect indicating that the damage to the device during hot carrier stress is localized at the drain end. The resulting reduction in drive current will cause a circuit to operate at a significantly slower speed than it was designed for and/or a race condition can be obtained which will affect circuit functionality. It is generally accepted that the degredation of N-MOSFET characteristics by negative charge build up during the maximum substrate current stress is due to interface trap generation which increases the level of captured hot electrons.

The mechanism by which the interface traps are formed is under debate. In one model [81] the interface traps are speculated to be formed by the bond breaking of Si-H which requires 0.3eV and for the resulting trivalent silicon dangling bond to becomes an electron trap. The resultant trapping of electrons with time causes the build up of negative charge at the silicon interface. The Si-H bonds are thought to occur during forming gas anneal and a hot electron would have to have an energy in excess of 3.5eV to surmount the Si-SiO_2 barrier and break the Si-H bond.

The second model used to describe the generation of interface traps [82] assumes that during the oxide growth, neutral trapping sites N_t^0 exist about 10nm from the SiO_2. Initially only hot holes can fill the neutral traps due to the electric field direction and a build up of positive trapped charge occurs, N_t^+. This positive charge then attracts the hot electrons and the original neutral traps N_{the}^0 are recovered as well as the formation of neutral interface states. The interface state is then charged negative N_{it}^-, during transconductance measurement or hot carrier stress at maximum substrate current conditions. Figure 5.8 shows the resulting evolution of the densities of the different trap sites with time.

The second model has also been extended recently [83,84,85] to describe the degredation of N-MOSFET's during low and high V_{GS} stress. During low gate voltage stress ($V_{GS} \approx V_{DS}/5$), the negatively charged interface traps are masked by trapped holes and there is no shift in the linear region device characteristics. Hot holes injected during low gate voltage stress can

produce neutral electron traps. Both of these effects can be revealed to significantly degrade the device characteristics if a short electron injection is performed at V_{GS}=V_{DS}. At high gate voltage stress corresponding to V_{GS}=V_{DS}, the damage is directly caused by electron trapping in the gate oxide from the hot electrons. The maximum device degredation occurs during maximum substrate current conditions although devices with poorly grown gate oxides or with lots of process-induced damage will degrade more by high gate voltage stresses.

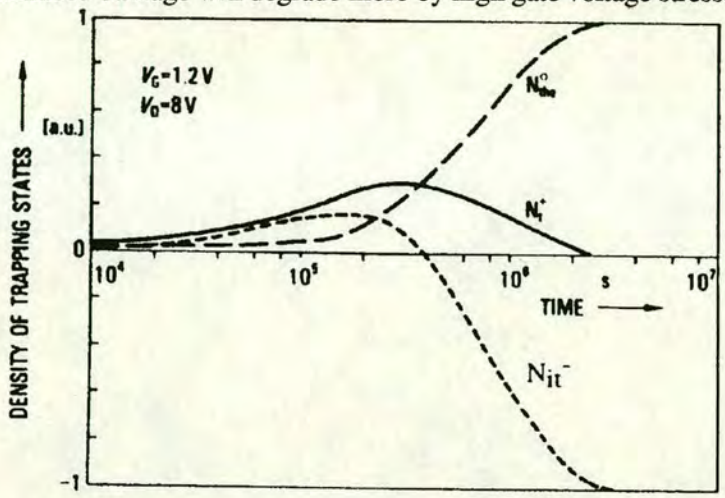


Figure 5.8 Evolution of the densities of N_t^+, N_{the}^0 and N_{it}^- with time [82].

Since there is no direct way of measuring the trapped charge densities, the substrate current is generally accepted as being a way to monitor hot carrier stress damage. While it can be appropriate to compare the substrate current between devices within a technology, it has been pointed out [81] that the substrate current does not give any indicator of the location of the maximum electric field and should therefore not be used to compare the hot carrier robustness of different technologies. Another problem with using the maximum substrate current as a monitor of hot carrier damage is that the same $I_{SUB(max)}$ can be obtained for different bias conditions which have obvious differences in the extent of device degradation.

The hot carrier degredation of similar devices with differing gate oxide thickness shows that thinner oxides are more robust than thicker ones [86]. This is thought to be due to the higher electric field over a thin oxide which repels hot electrons more strongly. This means that while the substrate current is higher in devices with thinner oxides, there are fewer hot electrons with sufficient energy required to cause damage to the device. This is another indication that using $I_{SUB(max)}$ as an indicator of hot carrier damage may be inaccurate.

The location of the damaged region is above or adjacent to the drain region depending on the location of the maximum electric field. The spread of the damaged region will differ

according to the depth of the maximum electric field, the distribution of defects at the Si-SiO_2 interface, the stress condition and stress duration. The length of the damaged region has been observed to be independent of device channel length so that the damaged region can become a large fraction of the device length in short channel devices [87,88].

5.1.5 Characterizing the Robustness of Devices to HCE

There is no general criterion to determine the point at which a MOS device has degraded from hot carriers such that it will affect circuit functionality. From the four parameters V_T, S, g_m, I_{dsat}, the drive current has been found to be the most sensitive device parameter to affect both circuit speed and functionality. In some cases, an arbitrary failure condition of say 5% degredation in drive current is used to define the lifetime at a specific stress condition. The lack of a method for predicting circuit lifetime can however, lead to a conservative approach of compromising transistor performance in order to improve transistor hot carrier lifetime. Circuit simulations can be done to determine what MOS device parameters shifts are most critical to circuit functionality and the failure criterion of a worst case transistor can be determined. Since circuits are usually designed for 10 years operating lifetime, accelerated stressing is required to assess the robustness to hot carrier effects in a reasonable time frame. An obvious way to accelerate device stress conditions is to use D.C. conditions and scale the lifetimes through the circuit duty cycle to A.C. conditions. In addition it is necessary to stress devices at higher drain voltages than the technology is designed for and to then extrapolate the lifetime back to the worst case operating voltages (to include the effects of ringing, etc.). It is worth noting here that the drain voltage should not be increased to the point that the MOS device will go into the Snap-Back condition as the stress mechanism will be different.

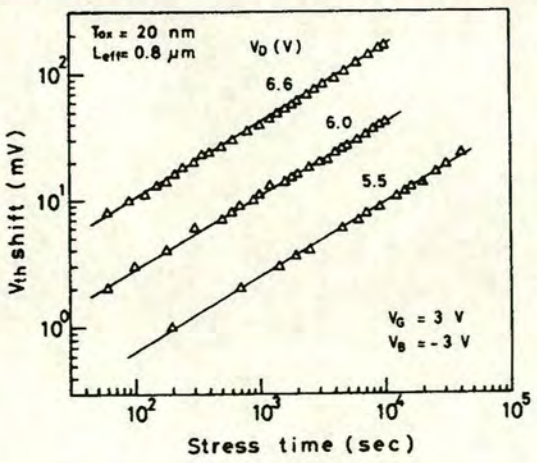


Figure 5.9 Logarithm of threshold voltage degredation verses logarithm of stress time showing the power law dependence [80].

The degredation in drain current over time has been found to obey the following power law dependence as is shown in Figure 5.9.

$$\Delta V_T = K \cdot t^p \qquad 5.7$$

where p is the gradient of the straight lines in Figure 5.9.

This type of plot can be used to find the accelerated lifetime of devices at different drain voltages. Figure 5.10 shows the correlation of substrate current with device lifetime τ which could be found from Figure 5.9, defined by a 10% reduction on drive current from stressing devices at the maximum substrate condition for high drain voltages. The slope of this plot m, is independent of technology and has been found to be around 3. Since I_{SUB} is proportional to the reciprocal of V_{DS} from Equation 5.6, Figure 5.11 can also be used to project the operating lifetime from accelerated stresses at high V_{DS} conditions. The lifetime of devices stressed under these conditions of interface state generation can be modelled as [85]

$$\tau_{N_{it}} = A \cdot L_{eff}^{r} \cdot \left(\frac{I_b}{W}\right)^{-m} \qquad 5.8$$

where r and A are fitting parameters for a given technology.

For the case of low gate voltage stress ($V_{GS} \approx V_{DS}/5$) which is associated with hole trapping [84], the device lifetime can be modelled by

$$\tau_{N_{ox,h}} = B \cdot L_{eff}^{r} \cdot \frac{\left(\frac{I_b}{I_d}\right)^{-n}}{\left(\frac{I_d}{W}\right)} \qquad 5.9$$

where n=8 and B, r are technology dependant fitting parameters.

The third type of stress [85] where electron traps are created when $V_{GS}=V_{DS}$,

$$\tau_{N_{ox,e}} = C \cdot L_{eff}^{r} \cdot \frac{\left(\frac{I_g}{I_d}\right)^{-l}}{\left(\frac{I_d}{W}\right)} \qquad 5.10$$

with l=1.9 and C,r technology dependent parameters.

In order to estimate the A.C. lifetime of a circuit [89], each of the three damage conditions from Equations 5.8-5.10 can be integrated over the A.C. circuit operating waveform to yield

the circuit lifetime τ_{AC} as follows

$$\frac{1}{\tau_{AC}} = \frac{1}{\tau_{N_{it}}} + \frac{1}{\tau_{N_{ox,h}}} + \frac{1}{\tau_{N_{ox,e}}} \qquad 5.11$$

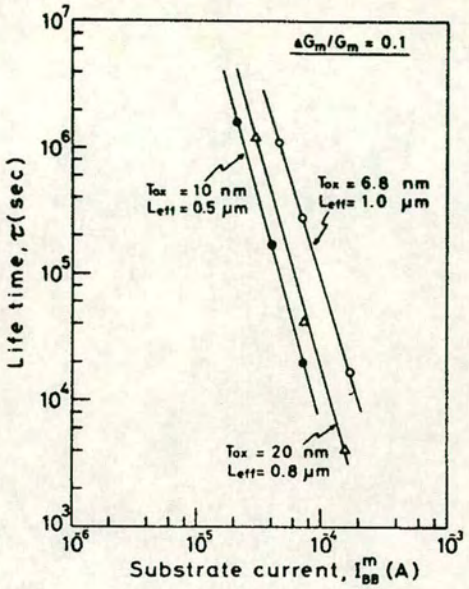


Figure 5.10 Plot of device lifetime verses substrate current which can be used to extrapolate accelerated hot carrier stress to operating conditions [80].

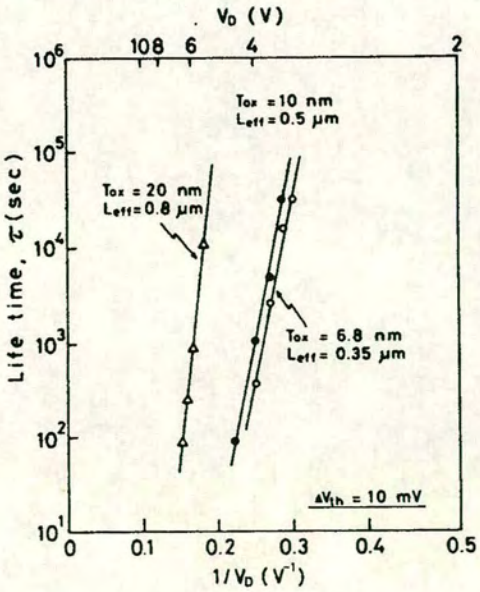


Figure 5.11 Plot of device lifetime verses the reciprocal of drain voltage which can be used to extrapolate accelerated hot carrier stress to operating conditions [80].

5.2 The Hot Carrier Effects in P-MOSFET Devices.

Hot carrier effects in PMOS devices are less severe than in NMOS devices primarily due to the 1 or 2 order reduction in impact ionization rate for a given electric field. In addition buried channel PMOS device have a current path which is well below the silicon surface which also reduced the susceptibility to hot carrier degredation. Hot carrier effects however do become important for surface channel PMOS devices with sub-half micron dimensions. The relatively high potential barrier for holes to cross inorder to reach the gate oxide means that only electrons resulting from impact ionization are able to cause damage to PMOS devices. The direction of the electric fields in a PMOS only favour electron injection into the gate oxide [90]. Most of the hot electrons travel through the oxide to form a gate electron current while some become trapped in the gate oxide and cause shifts in the device characteristics and result in an effective shortening of the channel length [91]. PMOS devices are therefore susceptible to hot electron induced punchthrough (HEIP) due to the reduction in the effective channel length to the point at which the source and drain short together [92]. The trapped electron charge causes an increase in both the transconductance and drain current but also a decrease in the threshold voltage. The worst case stress conditions are found when the gate current is a maximum which occurs around the threshold voltage of the device. At higher gate voltages, electron injection into the gate is suppressed because the gate becomes more repelling (negative). In a similar manner to NMOS devices the lifetime of PMOS devices operating at low voltages can be extrapolated through accelerated stress tests at higher drain voltages (more negative). Figure 5.12 shows the lifetime at maximum gate current stress conditions for an increase in drain current of 5% plotted against the reciprocal of drain voltage [92].

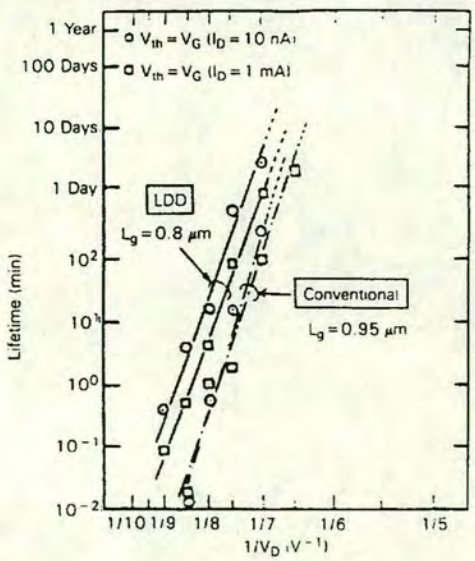


Figure 5.12 Device lifetime at maximum gate current plotted against reciprocal drain voltage.

5.3 Modifications to the Drain Region for CHE Robustness

5.3.1 Double Diffused Drain (DDD) Structure

The robustness of a MOS device to hot carrier effects can be increased by altering the drain profile. This is accomplished by grading the drain dopant such that a less abrupt diode is formed to the substrate and the maximum electric field is reduced by the voltage drop over the graded region. The penalty for the voltage drop is a reduced drive current which results in a performance/reliability trade-off.

The double diffused drain structure was implemented in N-MOSFET's as a way to reduce the maximum electric field in the channel as dimensions were shrunk to around 1.5μm without lowering the operating voltage of 5V. The DDD can provide a means of grading the drain through the implantation of both Arsenic and Phosphorous. Since phosphorous diffuses faster than arsenic a long drain dopant profile tail into the substrate can be achieved. Figure 5.13 shows a comparison of the dopant profile of a DDD process with a single diffused drain process. The dose of the arsenic is required to be high so that the resistance of the drain is low. Although this process is simple to implement, the long thermal drives required to achieve significant hot carrier lifetimes however, result in deep source/drain junctions which limit this method to 1.25μm channel lengths due to short channel effects [93].

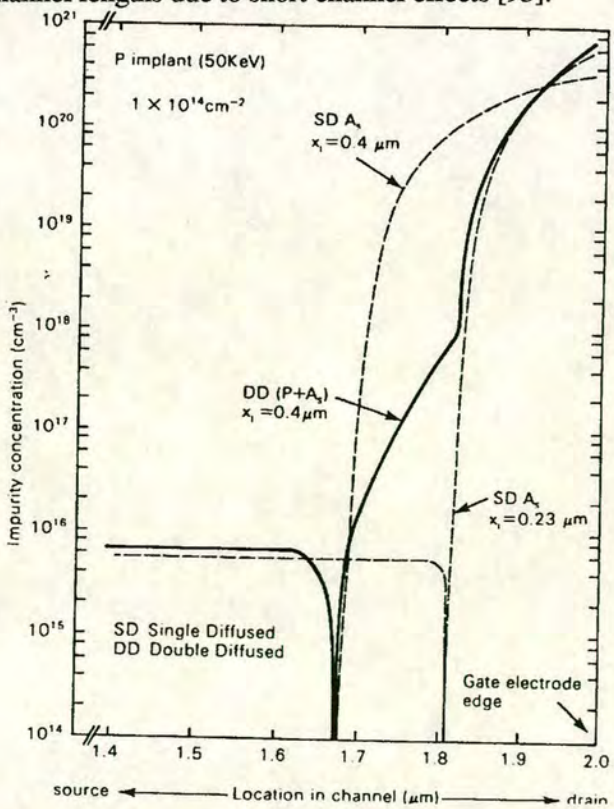


Figure 5.13 Lateral dopant profile from channel to drain regions for the case of a single - diffused drain and a double-diffused drain [94].

5.3.2 Lightly Doped Drain (LDD) Structure

The implementation of the lightly doped drain structure allowed N-MOS devices to be scaled to submicron dimensions at 5V operation with sufficient hot carrier reliability. Figure 5.14 shows the conventional LDD structure fabrication sequence as was first described by Ogura in 1980 [95]. The first implant is self aligned to the polysilicon gate and is relatively low dose (10^{12}-10^{13} cm^{-3}) in order to step the drain dopant concentration down thereby reducing the junction abruptness toward the lateral channel region. A CVD oxide is then deposited and etched anisotropically to form a spacer oxide over the polysilicon sidewall. The second implant is subsequently done with a dose of ~$10^{15}cm^{-3}$ which enables a low contact resistance to the drain to be achieved. In this way the length of the lightly doped region is defined by the oxide spacer width and the thermal drive cycle. This allows the scaling of the LDD region independently of the polysilicon gate length or heavy doped drain junction depth [96].

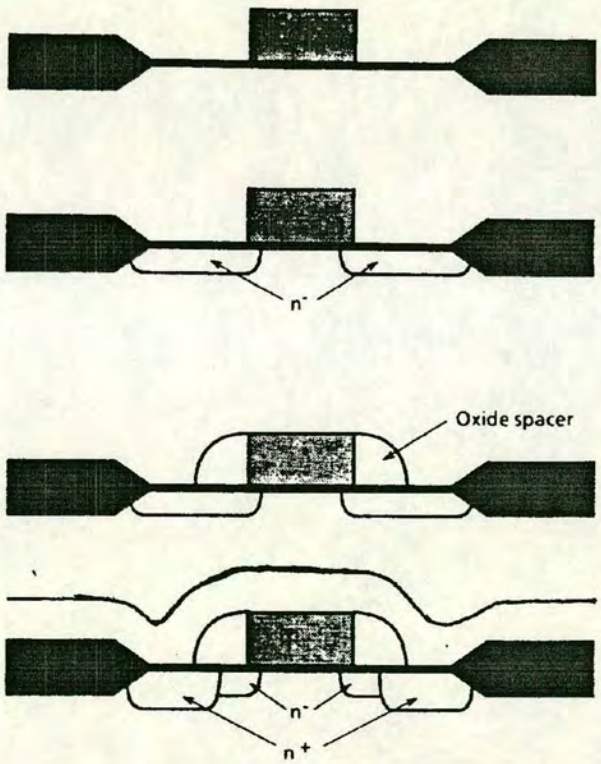


Figure 5.14 Conventional LDD fabrication sequence. (a) After Polysilicon patterning, a low dose implant is performed to form the lightly doped drain region. (b) A CVD oxide is deposited and then RIE etched to form oxide spacers around the gate. (c) The heavy dose implant required to lower the resistance of the drain and polysilicon gate are performed. (d) Silicided gates and source/drain areas are self aligned (SALICIDE) to lower the contact resistance further.

The use of an LDD has been shown to reduce the maximum electric field by 30-40% as shown in Figure 5.15. Figure 5.16 shows a cross-section of an LDD structure with associated physical attributes. The maximum electric field in an LDD structure has been obtained by modifying Equation 5.1, to give [81]

$$\varepsilon_{ymax}(LDD) = (V_{DS} - V_{DSSat})/(0.22 t_{ox}^{1/3} r_j^{1/3} + L_{n^-}) \qquad 5.12$$

where L_{n^-} is the length of the lightly doped drain region as shown in Figure 5.16. The above model however, predicts that the maximum electric field will be reduced further by increasing the length of the lightly doped or n^- region. In practice this is not the case and the maximum electric field actually starts to increase with increasing L_{n^-}. Another short coming is that the model does not account for the change in position of the maximum electric field as shown in Figure 5.15.

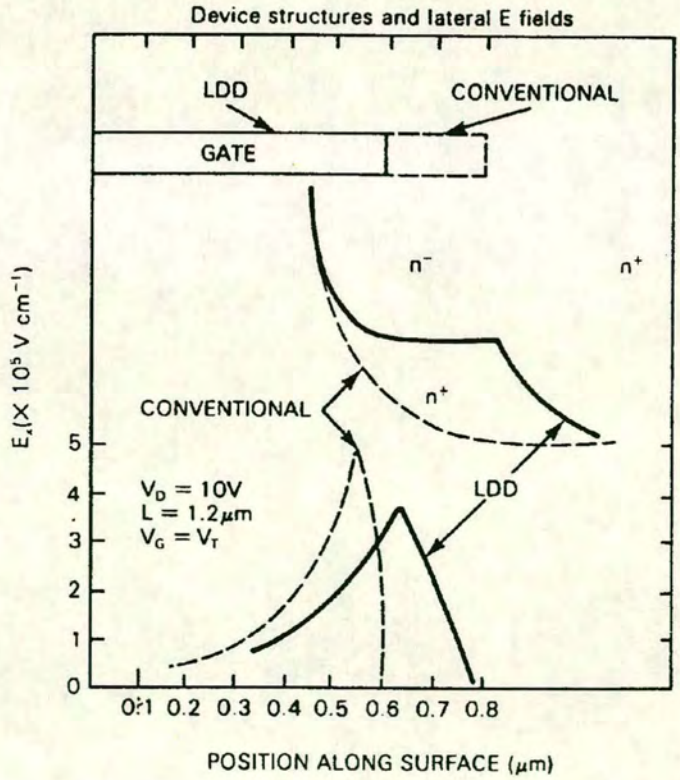


Figure 5.15 Reduction in the maximum electric field (and shift in position) at the silicon-silicon dioxide interface as a function of distance of an LDD structure compared to a single diffused drain structure [95].

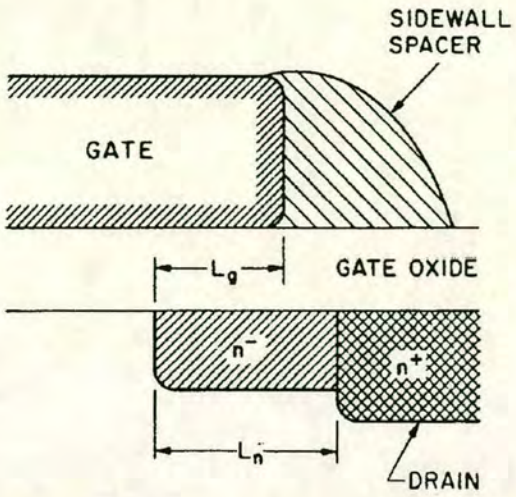


Figure 5.16 Schematic cross-section of a LDD structure.

Marayam et al [97] used a 2-D solution to Poisson's equation to determine that the maximum electric field is reduced most in a structure where the gate overlap of the drain is equal to the length of the lightly doped drain region. They also showed that the location of the peak electric field depends on the dose of the lightly doped implant. A lower dose implant was shown to provide better hot carrier lifetimes due to the peak electric field being located under the polysilicon gate. The condition when the gate overlap is small is referred to as weak overlap [98] and results in lower hot carrier lifetimes, attributed to the lack of control the gate field has over the hot carriers. This condition is best avoided and can occur from gate oxide bird's beak formation and off-axis LDD implant shadowing effects [99,100,101].

Figure 5.17 shows the reduction in drive current due to the increased resistance of the LDD which results in an additional voltage drop. The effect of this parasitic resistance is most pronounced in the linear region of operation where the effect of both drain and source resistance is accounted. During the drive current measurement, only the source resistance influences the measurement. The gains achievable in hot carrier robustness are therefore traded off with the reduction in device performance. The dose and length of the lightly doped region determine this trade off [102]. The exercise to determine the optimal LDD conditions for a given technology are however non-trivial. This has been termed as Drain Engineering.

Limitations to conventional LDD structures were due to too low LDD doses which resulted in large performance penalties and problems with weakly overlapped device control. Fully overlapped LDD devices gave better control but the thermal cycle required, limited device scaling to 0.7μm channel lengths from short channel effects.

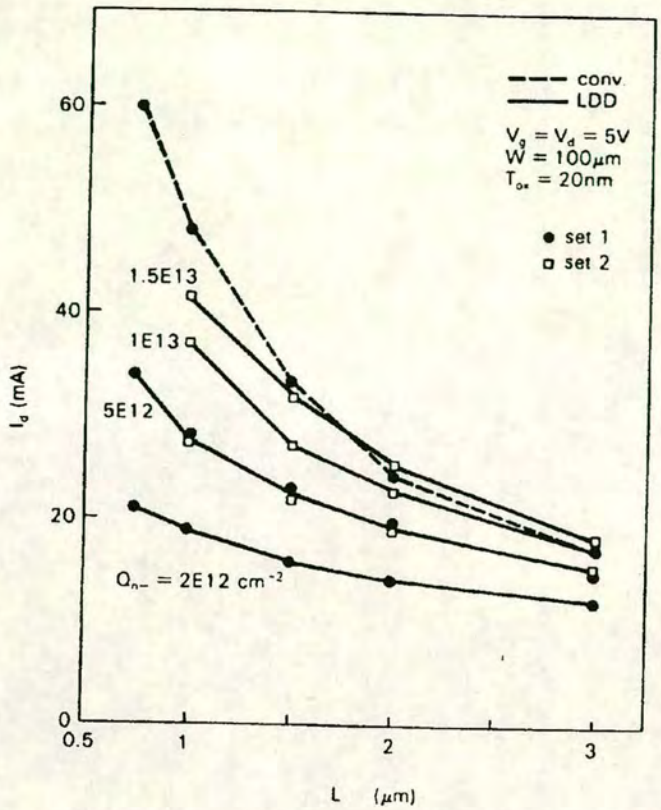


Figure 5.17 Reduction in saturation current for various channel lengths [103].

5.3.3 Improvements to the Conventional LDD Structure

In this section some of the improvements on the conventional LDD structure which have enabled device scaling below half micron channel lengths while maintaining hot carrier reliability will be summarized. It should be noted however that with the reduction in channel lengths below 0.5μm, there has also been a reduction in the operating voltage for the reason of reduced power consumption [6]. While this has resulted in reducing the hot carrier problem with device scaling, there has still been the need to provide more complex solutions to the hot carrier problem.

The first approach to modifying the conventional LDD structure is to increase the dose of the LDD implant (~10^{14} cm^{-3}) such that it becomes a Moderately Doped Drain (MDD) [104]. The increased dose results in the peak of the electric field being under the gate and the drive current degredation is reduced.

Another modification to the LDD structure is to implant the LDD implant deep such that the peak electric field is located deeper from the silicon surface [105]. This buried-LDD method can also be combined with a deep graded profile as shown in Figure 5.18 to provide better hot carrier lifetimes.

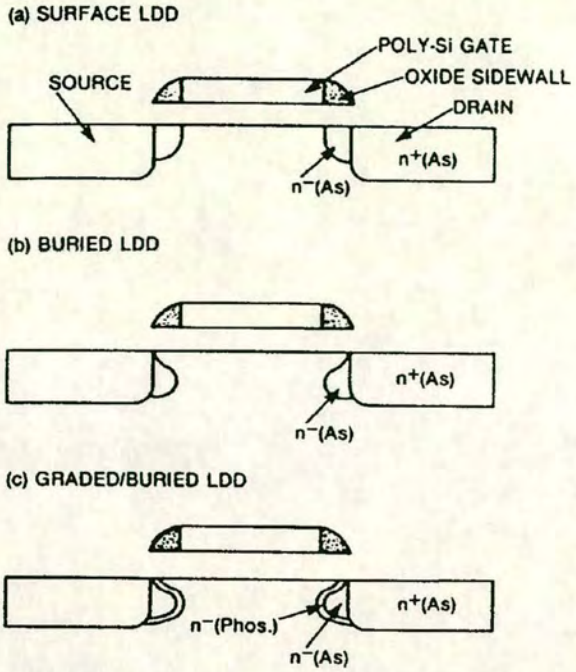


Figure 5.18 Schematic cross-sections of (a) conventional LDD, (b) Buried-LDD and (c) Graded Buried LDD [106].

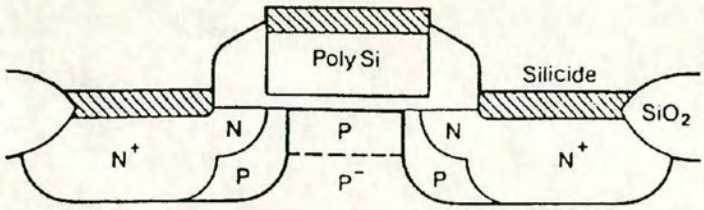


Figure 5.19 Halo doping structure as applied to a N-MOSFET [107].

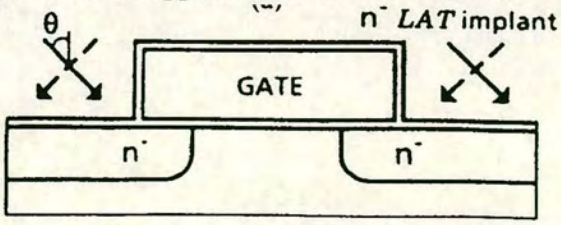


Figure 5.20 Large Angle Tilted Implanted Drain (LATID) structure [108].

The use of deep boron implants into the NMOS source/drain regions prior to the LDD process results in a Halo region around the source/drain regions [107]. This Halo region serves to increase the punchthrough voltage in the channel and increase the hot carrier resistance of the LDD device without degrading the drive current. A schematic is shown in Figure 5.19. The penalty incurred for this structure is increased sidewall junction capacitance.

Recently the self-aligned pocket implant (SPI) has been shown to reduce the extent of the increased sidewall capacitance [109].

The Inverse T-gate LDD (ITLDD) [110] and the Gate-Drain Overlapped LDD (GOLD) [111] structures and variations around them have been used to ensure the overlap of the gate over the LDD region. For both of these complex fabrication methods, the polysilicon gate is effectively extended over the LDD region while still maintaining the same electrical channel length as a conventional LDD.

The replacement of a CVD oxide spacer with a material with a larger dielectric constant has been shown to make partially overlapped devices behave more like fully overlapped devices. Materials such a silicon nitride and Ta_2O_5 have been considered [112,113]. The higher dielectric constant of the spacer enhances the gate fringing field and provides for better control of the gate over the n^- region. Problems associated with this technique are increased interface states at the spacer to oxide interface and the compatibility with a Self Aligned Silicide (SALICIDE) process.

A common method to increase the robustness of subhalf micron devices to hot carriers is to use Large Angle Tilted Implanted Drains (LATID) [114]. In this method the lower dose LDD implant is carried out at an angle between 45-60°. This enables the lateral and vertical dopant profiles to be optimized independently for hot carrier and short channel effects. The objective of this type of implant is to introduce the desired dopant concentration at the appropriate depth without the need to use a thermal drive to place it there. The wafer is also normally rotated and implanted in quads such that all four corners of the transistor are completely exposed to the implant beam. This results in negligible asymmetric effects from implant shadowing. One restriction in the use of this implant is that an implant at 37° with respect to the <100> silicon lattice should be avoided due to excessive channelling. Figure 5.20 shows the LATID technique. This technique is relatively simple to implement and is scalable due to its flexibility compared to the others mentioned in this section. Hori et al. [115] have shown this method to give adequate hot carrier reliability down to 0.25μm NMOS device lengths. Angled implantation for SPI and source/drain implants has also been shown to improve device operating characteristics symmetry.

An asymmetric design of a NMOS device has been shown to be optimized for both hot carrier effects and punchthrough resistance. The drain side of the device incorporated a LATID profile while the source side of the device uses a Halo doped implant. This device was shown to have good reliability and performance for a 0.25μm channel length technology [116]. In most applications however, the limiting device for performance and reliability is the pass

transistor which required to operate in both forward and reverse directions with symmetrical device parameters.

In this section, only N-MOSFET device have been considered so far. In PMOS devices the hot carrier effects only become a serious problem below 0.7μm channel lengths. For these smaller dimensions, the same structures and techniques can be applied to PMOS devices as were shown for NMOS devices.

5.4 Hardening Gate Oxides for Reduced CHE Degredation

5.4.1 Methods to Reduce the Hydrogen Content in Gate Oxides

Gate oxide dielectrics with low densities of trapping centres, interface state traps and fixed oxide charges are more resistant to hot carrier effects. It is also desirable to have oxides which contain a certain concentration of theses traps or charges, to exhibit minimal shift in device characteristics when stressed under hot carrier conditions. Hydrogen is either intentionally (forming gas alloy) or unintentionally (e.g. moisture absorption) incorporated at the Si-SiO_2 interface during processing where it passivates dangling bonds. Since the Si-H bond is relatively weak, hot carriers injected into the oxide can easily break these bonds with time to cause parametric shifts [117]. Hydrogen is thought to be incorporated during silicon processing from a variety of sources. PECVD silicon nitride passivation was shown to be a high source of hydrogen and as a result was replaced by PECVD silicon oxide [118]. Hydrogen anneals at high temperatures is another source of enhanced CHE. This resulted in the replacement of PSG for interlevel dielectrics by BPSG due to the fact that BPSG can be reflowed for planarization in nitrogen compared to the reflow of PSG in steam. Moisture from SOG and TEOS films have also been shown to increase the hot carrier degredation of MOS devices [119,120]. Gate oxides grown in a dry ambient have been shown to exhibit less hot carrier related effects compared to wet grown oxides. In a study by Ohji et al. [121], ultra-dry gate oxide growth showed improvements in CHE robustness over conventional dry oxide growth. In addition to the benefits of dry oxide growth for hot carrier effects, the charge to breakdown of dry oxides is significantly higher. It has been shown by Hsu et al. [122], that the amount of parametric shift during CHE can be reduced if the final process anneal is carrier out in a higher nitrogen to hydrogen containing ambient as shown in Figure 5.21. Another study showed an improvement if a nitrogen anneal is carried out after the hydrogen anneal [123].

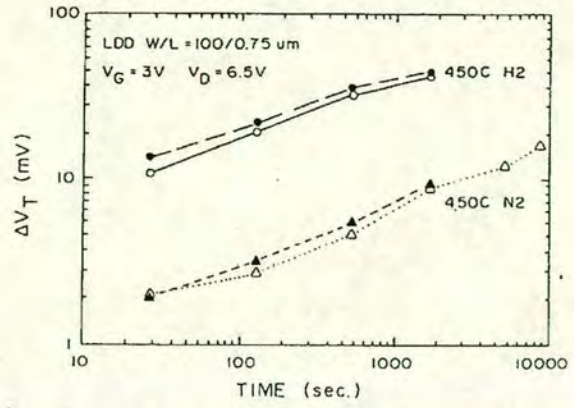


Figure 5.21 Reduced threshold voltage shift during hot carrier stressing with devices annealed in nitrogen compared to devices annealed in hydrogen [122].

5.4.2 CHE Degredation from Plasma Charging of Gate Oxides

Plasma charging of thin gate oxides during silicon processing can cause increased hot carrier effects as well as oxide breakdown degredation [124]. This is thought to be due primarily to the increase in electron trapped charge which occurs from the charging of the thin gate oxide during plasma etching, ashing and deposition (See Section 4.4). In a comprehensive statistical study by Mistry et al. [125], all three modes (N_{ss}, $N_{ox,h}$ $N_{ox,e}$) of hot carrier stress were examined to determine the affect plasma charging has on the hot carrier reliability of MOS devices. The result of this study showed that the most pronounced effect was an increase in $N_{ox,e}$ hot carrier stress in PMOS devices and to a lesser degree an effect on the NMOS device during electron trap creation. In a subsequent study by Li et al. [126], the degredation in CHE lifetime from the plasma damage to the oxide as a whole was differentiated from the plasma damage to the edge of a MOS device by measuring structures which enhanced each of the two modes of damage. The results of this study showed that the NMOS device was sensitive to only the edge damage from polysilicon overetching while the PMOS was sensitive to both electron trapping stress and the edge damage. This edge damage in the LDD region has also been shown to modulate the effective length of the device, effectively shortening the channel length and making the device more prone to CHE damage [127].

Improvements to the CHE robustness of MOS devices can therefore be accomplished by optimizing plasma processing equipment to reduce gate oxide charging from antenna effects. As with oxide degredation from plasma charging, design rules can be placed on antenna ratios such that the effect of plasma charging on hot carrier robustness can be effectively ignored.

5.4.3 Improved CHE Lifetimes from Halogen Incorporation

The presence of fluorine in gate oxides has been shown to harden oxides to hot carrier effects. Fluorine can be incorporated into the gate oxide during boron doping of a PMOS structure using BF_2+ source/drain implantation [128]. Increased MOS device lifetimes were also shown when fluorine was incorporated into the oxide by ion implantation [129]. A recent report [130] showed that hot carrier lifetimes were improved by the use of tungsten silicide due to the incorporation of fluorine from the WF_6 source gas. The mechanism for the improved reliability is thought to be the formation of Si-F bonds which are significantly stronger than Si-H bonds. Fluorine however has been shown to enhance the penetration of boron through the gate oxide to the silicon substrate in PMOS devices. The penetration of boron causes device instabilities and becomes a bigger problem as oxides are thinned. For this reason sources of fluorine are minimized in submicron device fabrication.

Chlorine can be incorporated into the gate oxide during thermal oxidation in TCA. Oxides grown by this method have shown improvements in hot carrier lifetime [131]. The mechanism for reduced device parameter shifts is a reduction in the amount of interface trapped charge by the formation of Si-Cl bonds during oxidation. The percentage of chlorine incorporated however is very low (<1%). With the reduction in thermal budget for shallow junctions, ion implantation has replaced POCl doping of polysilicon for NMOS devices eliminating this as a source of chlorine.

5.4.4 Incorporation of Nitrogen during Gate Oxide Growth

The incorporation of nitrogen into the gate oxide has shown a significant hardening of the oxide to hot carrier stressing conditions [132]. This is thought to be due to the termination of dangling bonds at the Si-SiO_2 interface by nitrogen bonds. The Si-N bonds are very strong compared to Si-H bonds and hence robust against hot carriers.

The first instance of thermal nitridation of silicon oxide was carried out by Ito et al [133]. NH_3 was used to nitride the thin oxide at atmospheric pressure and high temperatures of 900-1200°C. The nitrided oxide film resulted in a higher index of refraction compared to thermal oxide. An Auger emission spectroscopy (AES) study [134] showed that a large pile up of nitrogen occurred at both the oxide surface and the Si-SiO_2 interface which indicated that while the NH_3 reacted with the SiO_2, it also diffused very fast through the oxide to be incorporated at the silicon interface. Additional nitridation resulted in the bulk of the oxide being nitrided while the concentration at the oxide surface and silicon interface remaining essentially unchanged. While this form of nitridation reduces the level of interface trapped charge it also increases the levels of fixed charge and electron traps in the oxide [135,136].

Nitrided oxides grown by this method are therefore unsuitable for MOS devices due to inversion layer mobility reduction and $N_{ox,e}$ hot carrier damage.

The problems of increased electron traps with NH_3 nitridation were later shown to be alleviated with a reoxidation step after the nitridation. These oxides were termed re-oxidized nitrided oxides (RNO, ONO or ROXNOX) [137,138,139]. The thickness of nitrided oxide was noted to remain unchanged during the reoxidation step for short oxidation times. For excessive reoxidation times, the nitrogen that is piled at the silicon interface no longer provides a barrier to the oxidizing species. Hori et al.[140,141], proposed that the increase in electron traps associated with NH_3 nitridation occurred due to the hydrogen containing species resulting from the dissociation of ammonia. They also showed that the hydrogen concentration in the nitrided oxide reduced as a result of increased re-oxidation.

Thermal re-oxidation of nitrided oxides using the method just discussed may not reduce the level of electron traps enough. In addition, the fixed charge density remains high in these films. One way to reduce these problems is to perform the nitridation at low pressure or to use rapid thermal processing (RTP) nitridation [142]. Low pressure and RTP nitridation are thought to bring about a slower increase, peak and then decrease in the levels of fixed charge, interface trapped charge and electron trapped charge with increasing nitrogen content. There is therefore more control over the reduction in these charges from the re-oxidation step. This also means that relatively light nitridations with re-oxidations can be used to optimize the device characteristics in the pre-turnaround region. In the low pressure NH_3 nitridation process proposed by Gross et al. [143], the re-oxidation step removed the nitrogen incorporated into the bulk of the oxide to lower the fixed charge density without reducing the 8 atomic percent of nitrogen at the Si-SiO_2 interface. In a complete study by Momose et al. [144], the optimum nitrogen concentration at the silicon interface was found to be in the range of 0.2-1 atomic percent. Since different nitridation and re-oxidation conditions result in a different set of charges in the oxide, the aim of the re-oxidation step should be to reduce the level of fixed charge to that which is slightly higher than conventional oxide. This is because excess re-oxidation results in reducing the immunity of the ROXNOX to hot carriers due to a reduced nitrogen concentration at the silicon interface.

In order to reduce the density of electron traps with ROXNOX oxides which predominantly effects PMOSFET hot carrier lifetimes, the incorporation of hydrogen from ammonia is eliminated by the use of N_2O nitridation of silicon [145]. Figure 5.22 shows Secondary Ion Mass Spectrometry (SIMS) depth profiles through different oxide films [146]. The hydrogen concentration can be seen to be significantly reduced with nitridation in N_2O compared to

NH_3. The growth characteristics for N_2O nitridation of silicon [147] are shown in Figure 5.23. The growth was determined to be limited by the diffusion through the bulk of the oxide rather than through the piled up nitrogen at the silicon interface and shows a square root growth rate dependence. MOS devices fabricated with N_2O grown oxides showed reduced damage from interface traps in NMOS devices and reduced damage from electron trap creation in PMOS devices during hot carrier stressing albeit the interface trap levels before stress were higher than conventional oxides [148]. Figure 5.24 shows the effect of the interface traps on the mobility of electrons in NMOS devices. Although the mobility is reduced at low gate biases, the mobility is higher in the inversion region with these types of nitrided oxides [149]. The main problem with N_2O nitridation of silicon, however is the slow oxide growth rate. This lead to the development of N_2O nitridation of silicon oxide which resulted in a wide range of oxide thicknesses with a lower thermal budget [150]. After the oxide was grown in a dry O_2 ambient, the N_2O nitridation was carried out at 800-850°C. The nitridation step resulted in a 35Å increase to grow a 85Å oxynitride film. Additional re-oxidation of N_2O nitrided oxides [151] surprisingly showed that the nitrogen peak was displaced from the silicon interface by a newly grown oxide layer. There would therefore seem to be no benefit in doing the re-oxidation step for this case.

There have been various studies of the oxynitride growth characteristics in N_2O which do not corroborate [152,153,154]. The growth rate however is accepted to be limited by the surface reaction rate. In a paper by Tobin et al. [155], these discrepancies were attributed to the concentration of NO species which are responsible for the nitridation reaction in N_2O ambients. In a later study [156], conventional oxides were nitrided in NO or N_2O. The NO process was shown to incorporate an equivalent nitrogen concentration as a N_2O process but at a significantly lower temperature whilst maintaining a similar level of hot carrier robustness. Maiti et al. [157], subsequently showed significant improvements to the gate oxide breakdown at the field oxide edge by incorporating a re-oxidation of a NO nitrided oxide.

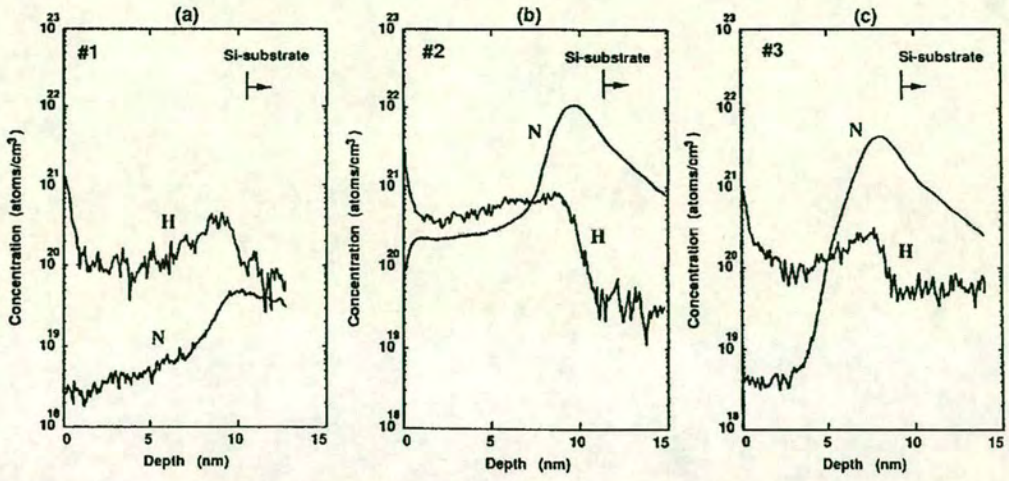


Figure 5.22 SIMS depth profiles of (a) conventional oxide, (b) NH_3-nitrided oxide and (c) N_2O-nitrided silicon films [146].

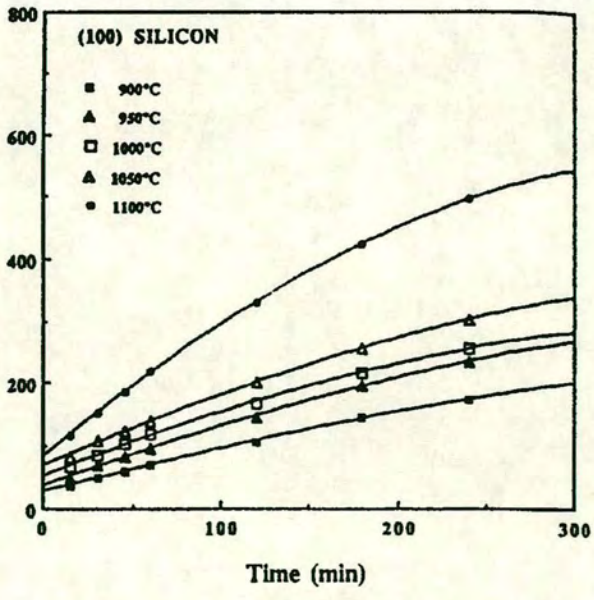


Figure 5.23 Growth rate dependence of silicon oxynitridation in N_2O [147].

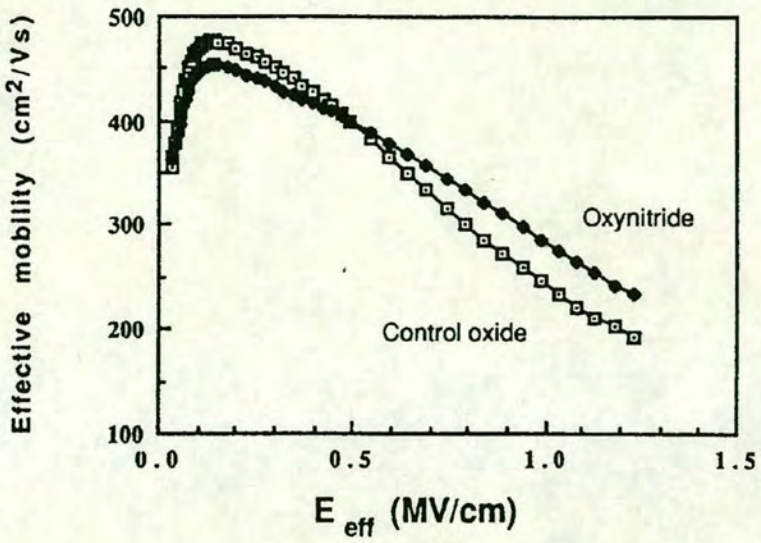


Figure 5.24 Comparison of low field and inversion layer mobility in N_2O nitrided oxide with conventional oxide films [149].

5.4.5 Incorporation of Nitrogen after Gate Oxide Growth

Nitrogen can be placed at the Si-SiO_2 interface by other methods. One such method first proposed by Haddad et al. [158], is to implant nitrogen through the polysilicon gate material where it piles at both the polysilicon-silicon oxide and the Si-SiO_2 interface by diffusion. This method was used by another group [159] to show improvements over conventional oxide in the hot carrier robustness of thin oxides. The penetration of boron through the oxide to the silicon crystal was reduced even in the presence of fluorine by the nitrogen peak at the polysilicon-silicon oxide interface [160]. The same group [161] have also shown improvements in hot carrier robustness by implanting nitrogen into the source/drain areas after the gate is patterned. Apparently the nitrogen piles-up to the silicon surface below the spacer oxide in the LDD region. A recent method [162] of incorporating nitrogen into the gate oxide is by including nitrogen during polysilicon gate deposition by rapid thermal CVD. A capping layer of undoped polysilicon is then deposited. Devices incorporating this technique showed reduced boron penetration and improved reliability.

CHAPTER 6

Characterisation of Silicon Oxidation Inhibition from Low Dose Nitrogen Implantation

The oxidation growth kinetics for low dose molecular nitrogen implanted silicon are extensively studied in this chapter in order to determine the degree of oxidation retardation for the application to mixed gate oxide thickness circuits. This work is also required to select different implant conditions to assess the electrical results of MOS devices with the same oxide thickness grown with different contents of nitrogen in the silicon.

6.1 Background

Nitrogen implantation into silicon has been previously studied by W.J.M.J. Josquin and Y. Tamminga [163,164]. In both these studies, relatively high nitrogen doses of 10^{15}-10^{16} atoms/cm^2 were used. In the first study [163] the initial N_2^+ implant redistribution upon annealing, showed two distinct peaks of N concentration. The first peak was seen to occur at the silicon/native oxide interface and the second peak was located where the amorphous silicon region met the crystalline silicon which corresponded to the implant range. The depth of the amorphized silicon was seen to increase with dose. For high nitrogen doses, the amorphous silicon region was not recrystallized until extensive annealing at 1000° C. Thermal annealing in an inert ambient showed the transfer of nitrogen atoms from the peak located in the bulk of the silicon to the peak at the silicon surface. The increased nitrogen diffusion to the silicon surface however, saturates at 4e21 atoms/cm^3 or ~8% [N] in silicon for a nitrogen implant dose of 10^{16} atoms/cm^2. The excess nitrogen was shown to be evaporated from the silicon surface [164] in an inert ambient. Increasing the nitrogen dose in the silicon did increase the nitrogen peak concentration at both the silicon surface and in the bulk silicon. Wafers which were implanted with nitrogen through a screen oxide layer also showed the concentration of the two peaks to increase with nitrogen dose. When bare silicon wafers were implanted with nitrogen and subsequently oxidized, a dependency on the oxidizing ambient conditions was found. For dry oxidations, the nitrogen was seen to diffuse from the bulk peak so that the whole dose of nitrogen in the silicon accumulates at the silicon surface and inhibits the oxidation rate. The re-crystallization of the silicon in the bulk now occurs in the absence of nitrogen. It was shown in [164] that the nitrogen concentration at the silicon surface has to be reduced (by evaporation from the silicon surface) to a concentration low enough to allow oxidation to proceed. This critical concentration of nitrogen in silicon

can be interpreted into an effective thickness of Si_3N_4 which needs to be consumed by oxidation before normal Deal-Grove oxidation characteristics of silicon will occur. The oxidation kinetics of silicon nitride have been studied by Enomoto et al. [165] to show a parabolic silicon oxide growth rate as shown in Equation 6.1. The thickness of silicon nitride that has been converted into silicon oxide is denoted by x_N, C is the rate constant, t is the oxidation time and τ_1 is the oxidation time adjustment to account for an initial oxide that was grown on the silicon nitride.

$$x_N = C(t + \tau_1)^{2/3} \qquad 6.1$$

Silicon oxidizes approximately 25 times faster than silicon nitride. Increasing the dose in this case, thickens the effective silicon nitride. For the case of wet oxidation, the nitrogen from the bulk peak in the silicon does not diffuse fast enough so that only the initial peak concentration at the silicon surface inhibits the oxidation. The effective silicon nitride thickness is thinner for wet oxidations compared to dry oxidations. The oxidation resistance is determined by a competition between the loss of nitrogen due to oxidation of the nitride-like layer and the buildup of a new nitride-like layer by nitrogen that diffuses from the bulk of the silicon toward the Si/SiO_2 interface [163].

6.2 Experimentation to Determine if Nitrogen Diffuses Through Silicon Oxide

The experimentation by Josquin and Tamminga [163] did not reveal the diffusion rate of nitrogen through silicon oxide. The diffusion of nitrogen through the oxide during oxidation would have a direct influence on the oxide growth characteristics such that the rate limiting step could change from reaction rate to diffusion rate limited. This section investigates the behaviour of nitrogen in the oxide, since the shape of the oxidation characteristics reported in this chapter could depend on oxidation conditions such as partial pressure or oxidation temperature.

In this study, silicon <100> wafers with 20-40 Ω-cm boron doping, were implanted with 10^{15} N_2^+ atoms/cm^2 at 100 KeV with 7° tilt angle. The implants were performed using an Eaton 6200 medium current ion implanter with a N_2 gas source. The N_2^+ beam current was 50μA. The wafers were then annealed at 975° C for 0, 30, 60, 180 or 360 seconds using rapid thermal annealing (RTA). Secondary ion mass spectroscopy (SIMS) was used to profile the nitrogen concentration. C_s^+ primary ions with a detection area of 15μm diameter were

optimized for sensitivity to molecular SiN. The samples were coated with a thin gold layer prior to analysis and additional electron gun bombardment was utilized for charge compensation of the oxide during the SIMS profiles. Conversion of the impurity ion counts to concentrations was achieved by using a relative sensitivity factor (RSF) derived from nitrogen implanted silicon reference material. The depth scales were calibrated by measurement of the total crater depths on a Tencor Alpha-step 200RD profileometer.

Figure 6.1 shows the six nitrogen profiles in silicon oxide and silicon. The as-implanted N profile shows the expected Gaussian profile. The nitrogen profile is distorted at the silicon oxide-silicon interface due to the primary ion sputtering rate deviation. The RTA nitrogen profiles show the two peaks described in [163]. For this dose of 10^{15} atoms/cm^2, the nitrogen concentration at the silicon oxide-silicon interface is seen to be approximately 6%. With increased anneal time, the N peak in the bulk of the silicon is shown to move deeper into the silicon as the nitrogen is supplied to increase the concentration of the N peak at the silicon oxide-silicon interface. This behaviour is in agreement with [163]. These nitrogen profiles in Figure 6.1 however, also show that there is no diffusion of nitrogen through the oxide. The nitrogen that is located in the oxide as a result of the ion implantation, is redistributed such that there is a relatively low concentration pile-up at the oxide surface. This relatively low concentration of nitrogen in the oxide should have little effect on the diffusing oxidant species [166]. The conclusion of this study is that the oxidation growth characteristics of nitrogen implanted silicon will be determined by the oxidation rate of the silicon nitride-like layer at the SiO_2-Si interface and hence the oxidation rate will be reaction rate limited.

6.3 Characterisation of Silicon Oxidation Under Low Dose N_2^+ Implantation

6.3.1 Introduction

The use of low dose nitrogen implantation into silicon has been proposed to retard the oxidation rate of silicon, allowing the simultaneous growth of different thicknesses of gate oxides on the same silicon wafer [10]. In that study, two process flows were used to implement the nitrogen implant. The following lists the process steps in order.

1. Implantation of N_2^+ into Si through a sacrificial or screen oxide.
2. A high temperature anneal to redistribute the nitrogen and reduce the implant damage.
3. Removal of the sacrificial oxide.
4. Growth of the gate oxide.

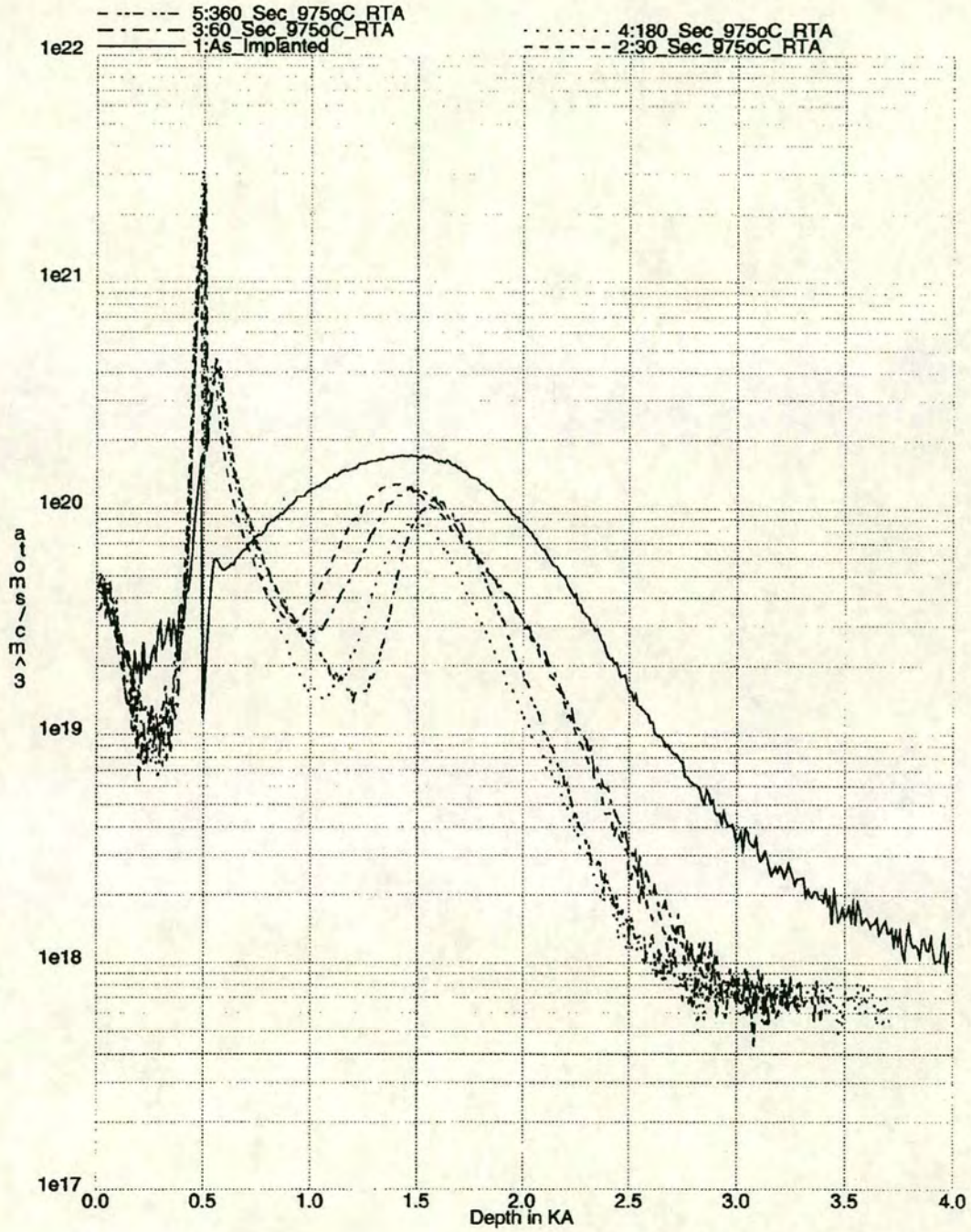


Figure 6.1 SIMS profiles of 10^{15} N_2^+ atoms/cm^2 ion implant at 100KeV in silicon through a 540Å screen oxide for (a) as-implanted, (b) 30 second RTA, (c) 60 second RTA, (d) 180 second RTA and (e) 360 second RTA at 975° C.

In the work of Soleimani et al. [10], wafers were processed using this flow with and without

the high temperature anneal step. While that work showed the retardation of the oxidation rate of silicon for both process flows, it was only a feasibility study and did not investigate the oxidation for different processing conditions such as nitrogen implant dose, nitrogen implant energy or oxidation time. The purpose of the following work is to characterize the silicon oxidation rate for different implant conditions for the case of the process flow without the high temperature anneal. This process flow was chosen for several reasons. Firstly, the elimination of the high temperature step enabled the nitrogen implant technique to be easily integrated into a 0.5μm CMOS manufacturing process. The inclusion of the high temperature anneal step would require some modifications to the transistor implant profiles to reduce short channel effects. Secondly, in the study by Josquin and Tamminga [163], the ion implant damage was only completely removed after extensive annealing.

6.3.2 Description of the Experimental Conditions

The silicon wafers used in this work were 150mm, boron doped <100> Czochralski crystal with resistivities ranging from 20 to 40 Ω-cm. Ellipsometer measurements of a 15Å native oxide corresponded to a physical thickness of 6-8Å. The wafers were cleaned using a conventional RCA method and a 225Å dry oxide was grown at 900°C to serve as a screen oxide. The molecular nitrogen or N_2^+ ion implants were done using an Eaton 6200 medium current implanter as described in Section 6.2. Four wafers were implanted with nitrogen for each of the conditions listed in Table 6.1. After ion implantation the screen oxide was removed by etching in 10:1 buffered hydrofluoric acid for 50 seconds. The samples were then oxidised at atmospheric pressure for 3, 5, 8 or 12 minutes in dry O_2 at 900°C for each of the implant conditions in Table 6.1. In order to provide a control, some wafers were oxidised for 3, 5, 8, or 12 minutes without any nitrogen implant. After oxidation, the wafers were annealed at 900°C in N_2 for 30 minutes to reduce the fixed oxide charge. The wafers were loaded in and out of the conventional furnace in N_2 at 800°C. The temperature ramps were also performed in inert N_2. Figure 6.2 shows the process steps. The dielectric thickness was measured using standard ellipsometry. The dielectric constant for these oxides were assumed to be that of conventional silicon oxide from earlier work which correlated measured ellipometer thicknesses with transmission electron micrographs (TEM) of oxides with a high concentration of nitrogen [166]. A nine point average was used to represent the oxide thickness on a wafer. For nitrogen doses less than $2x10^{15}$ cm^{-2} the intra-wafer uniformities were 3Å, and 8Å for the highest dose.

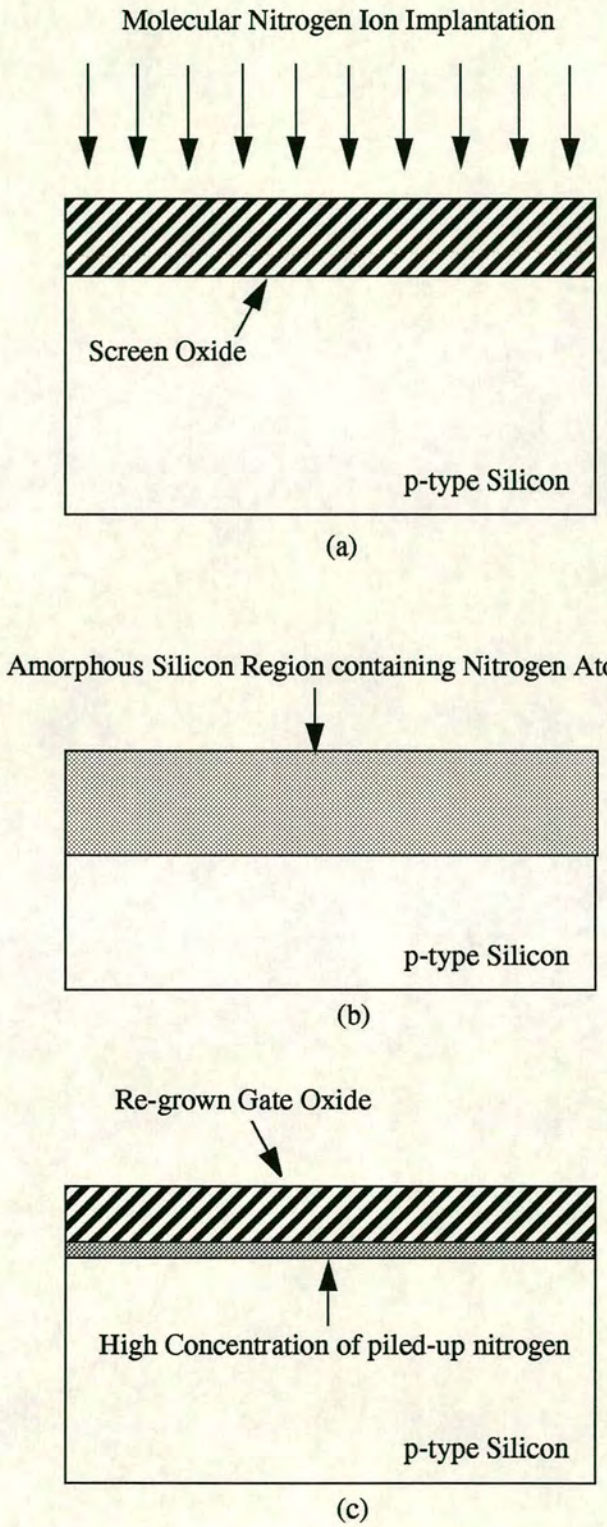


Figure 6.2 Process sequence for the oxidation of molecular nitrogen implanted silicon (a) Implantation of N_2^+ through a screen oxide. (b) The sacrificial oxide is stripped using 10:1 HF. (c) The gate oxide is grown in the presence of piled-up nitrogen at the silicon surface.

6.3.3 Oxidation Characteristics of Low Dose N_2^+ Implanted Silicon

The measured nine point average oxide thickness value for the conditions listed in Table 6.1 and the control wafers are listed in Table 6.2. Figure 6.3 shows this data plotted for the 10 KeV implant energy only. The dependency of oxide thickness with nitrogen dose starts at around 4e14 cm^{-2}. The oxide thickness below this dose is practically the same as that for the control wafers. For doses higher than 4e14 cm^{-2}, the oxidation rate falls off quite sharply. Figure 6.4 and 6.5 show the oxide growth data for the 20 and 30 KeV nitrogen implants respectively. The oxidation rate is retarded at even lower doses of approximately 1e14 cm^{-2} for the higher implant energies. For nitrogen doses higher than 1e15 cm^{-2}, the oxidation is completely inhibited. This behaviour can be more clearly seen in Figures 6.6-6.8 which re-plot the measured oxidation thickness against oxidation time for each of the implant energies. The oxidation growth thickness can be seen to flatten out as the oxidation proceeds for the higher doses. The 10 KeV samples show less dependence of oxide thickness with nitrogen dose due to the fact that the peak of the implant is located within the screen oxide and results in a loss of the major portion of the N_2 dose when the screen oxide is stripped off. When the oxidation curves from Figures 6.3-6.5 are overlaid, there can be seen to be a common curve for each oxidation time which is offset in dose due to the implant energy. This suggests that the retardation of the oxidation is due to the total dose of N_2 that is implanted into the silicon and then redistributed at the silicon surface. In order to determine when the oxidation would proceed in the high dose implanted wafers, wafers with the higher doses of 1e15 and 2e15 cm^{-2} were continually reoxidised and re-measured. Figure 6.9 shows the measured oxide thickness plotted against the reoxidation time. The oxidation is shown to proceed in the order of the implanted nitrogen dose in the silicon. For the same implant dose, the lower energy shows a quicker return to normal oxidation characteristics. The nitrogen concentration at the silicon surface has to be reduced to a low enough level for oxidation to proceed. The point at which oxidation proceeds in Figure 6.9 can be considered to be the point at which the effective thickness of silicon nitride has been reduced to zero. Even after 4 hours of oxidation in dry O_2 at 900° C, the 2e15 cm^{-2} 30 KeV N_2 wafers show no oxidation.

For practical purposes the useful range of molecular nitrogen implant doses is in the range of 1e14cm^{-2} to 1e15cm^{-2} since the oxidation time of the gate oxide is around 10 minutes. Also, since the N_2^+ beam current on the implanter is around 50μA, a dose of 1e15 cm^{-2} would take approximately 4 minutes, which is relatively long for a manufacturing process. It

is desirable to implant the nitrogen peak beyond the silicon oxide and into the silicon to control the uniformity of the dose better. High nitrogen implant energies should also be avoided so that the depth of the amorphous silicon region is minimised.

Energy (KeV)	N_2^+ Dose (cm^{-2})				
10	1e14	2e14	4e14	1e15	2e15
20	1e14	2e14	4e14	1e15	2e15
30	1e14	2e14	4e14	1e15	2e15

Table 6.1 Ion implant conditions for the experiment.

N_2^+ Dose	Energy (KeV	Oxide Thickness (Å)for Oxidation Times of			
		3 mins	5 mins	8 mins	12 mins
Control		64.9	72.7	81.7	93.0
1e14	10	64.0	71.5	80.4	91.1
2e14	10	62.9	70.2	78.3	89.2
4e14	10	60.4	67.3	74.4	88.1
1e15	10	49.2	58.4	65.8	74.6
2e15	10	36.5	39.5	46.2	51.7
1e14	20	62.2	67.2	75.3	89.3
2e14	20	53.0	60.7	69.9	78.3
4e14	20	38.7	43.5	50.2	57.4
1e15	20	22.1	25.5	25.5	24.5
2e15	20	18.2	21.7	23.0	19.7
1e14	30	59.8	65.4	73.4	85.4
2e14	30	48.4	55.0	62.1	71.7
4e14	30	31.9	34.9	40.0	43.7
1e15	30	19.3	23.4	20.1	21.1
2e15	30	18.9	22.1	18.5	20.1

Table 6.2 Measured oxide thickness for wafers implanted with molecular nitrogen at different conditions and oxidised for different times in dry O_2 at 900° C.

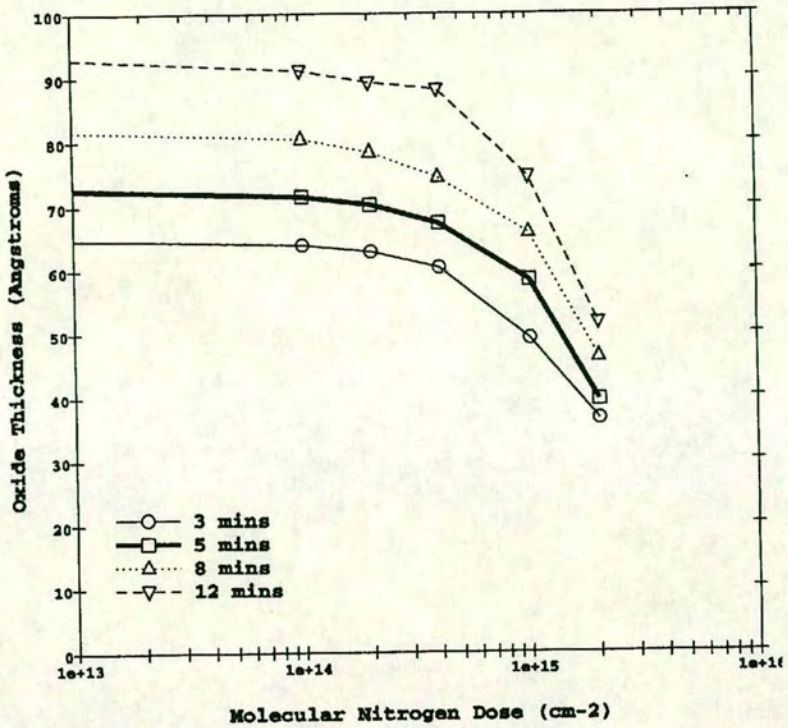


Figure 6.3 Measured oxide thickness verses molecular nitrogen implant dose at 10KeV implant energy for different oxidation times.

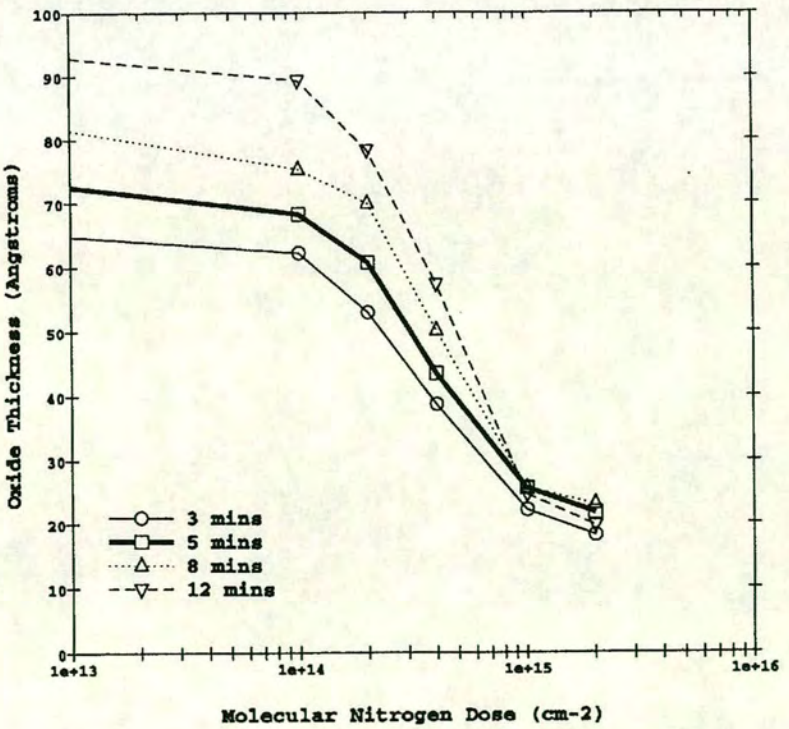


Figure 6.4 Measured oxide thickness verses molecular nitrogen implant dose at 20KeV implant energy for different oxidation times.

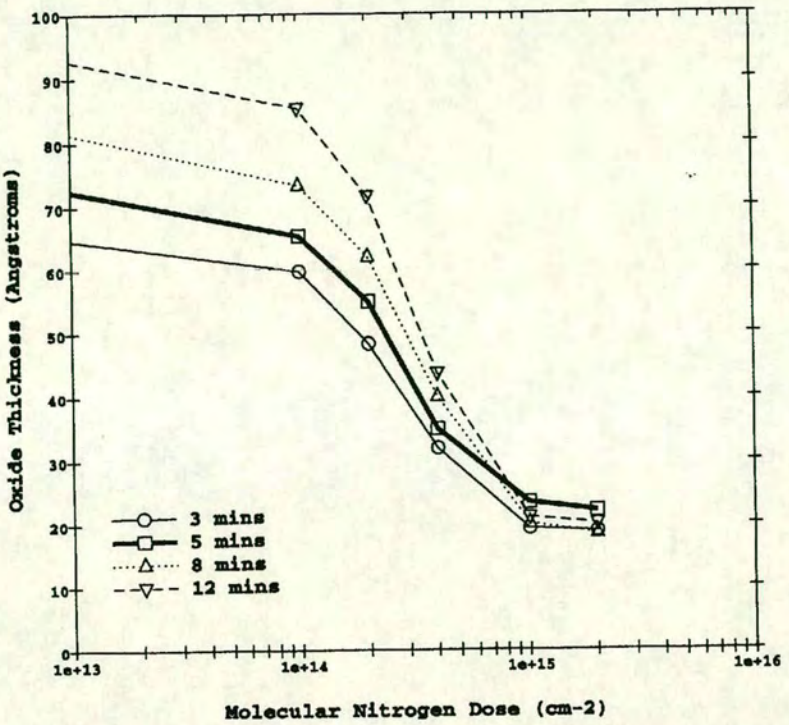


Figure 6.5 Measured oxide thickness verses molecular nitrogen implant dose at 30KeV implant energy for different oxidation times.

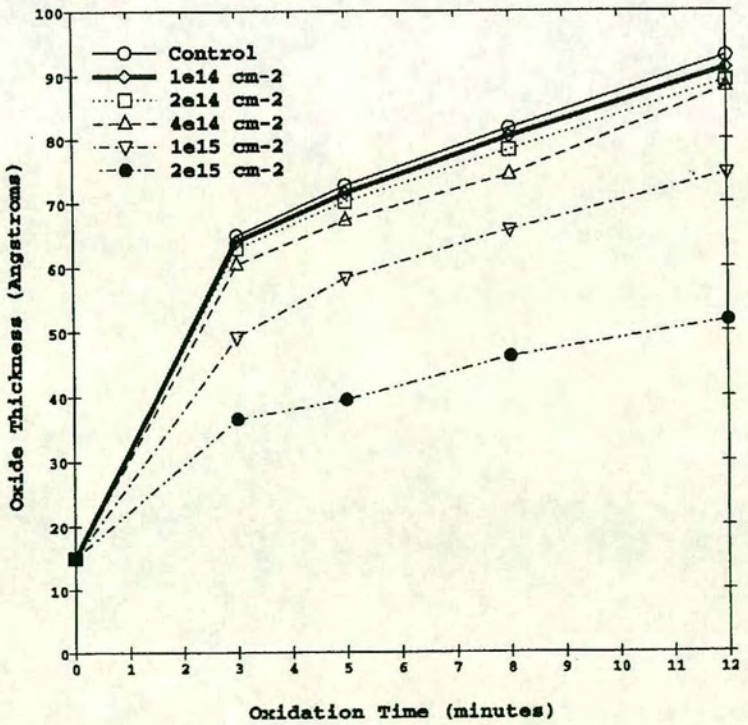


Figure 6.6 Measured oxide thickness verses oxidation time for different nitrogen implant doses at 10KeV implant energy.

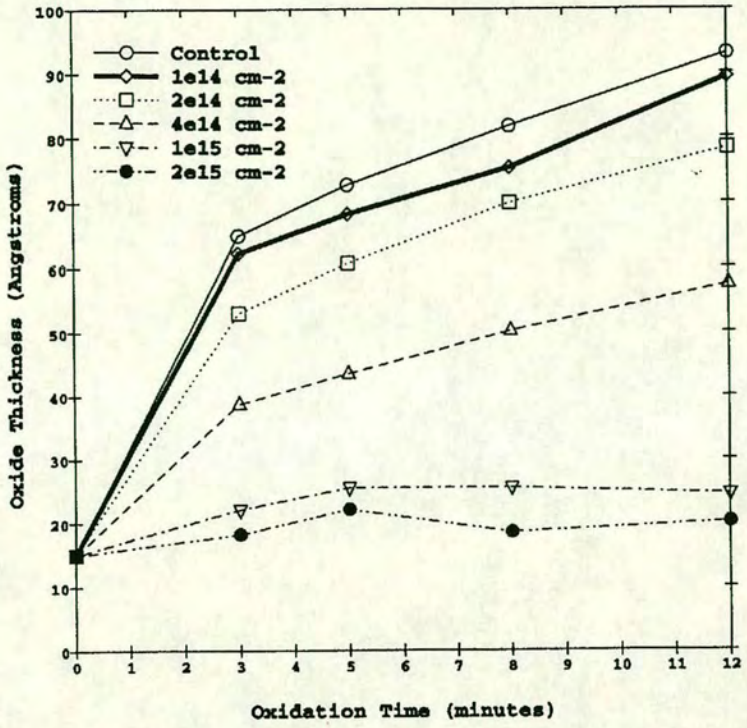


Figure 6.7 Measured oxide thickness verses oxidation time for different nitrogen implant doses at 20KeV implant energy.

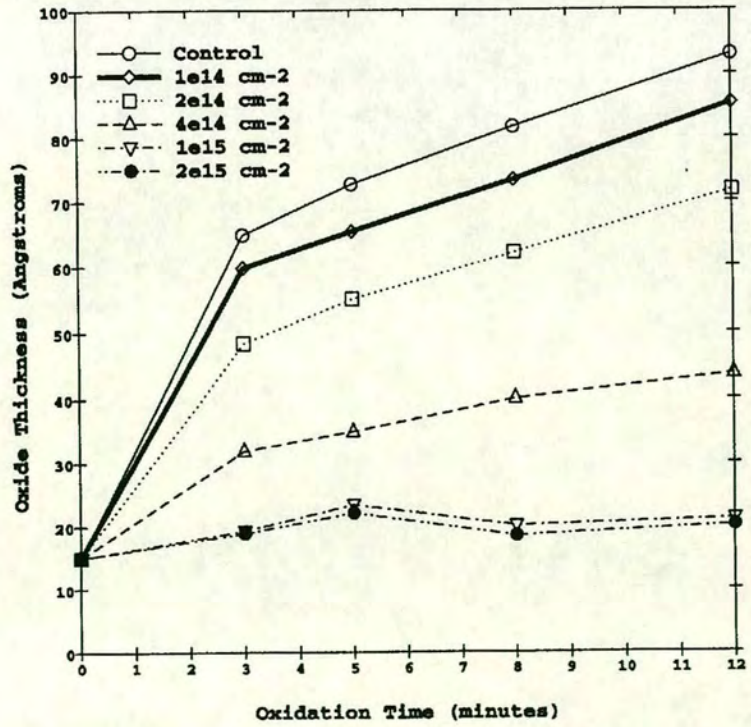


Figure 6.8 Measured oxide thickness verses oxidation time for different nitrogen implant doses at 30KeV implant energy.

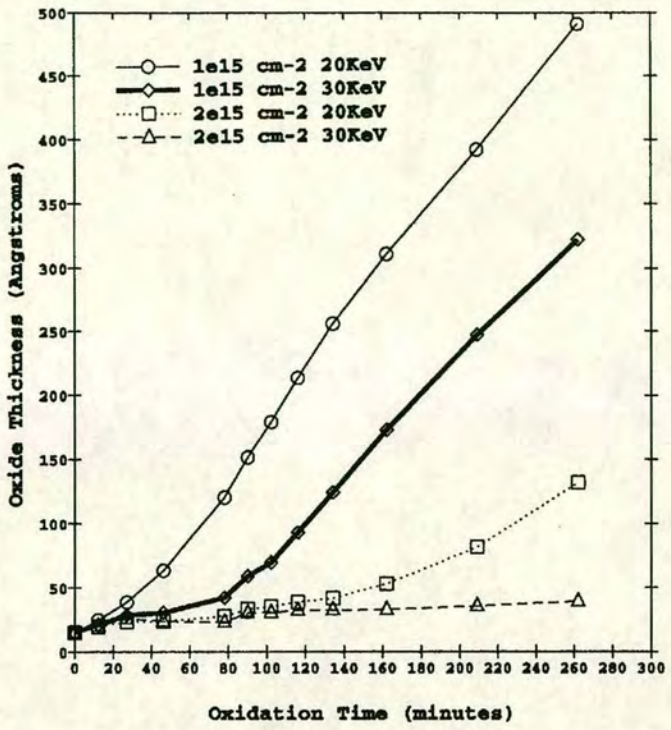


Figure 6.9 Measured oxide thickness verses time for the high dose molecular nitrogen implants after several reoxidation steps at 900° C dry O_2.

6.4 Conclusions

The oxidation characteristics of low dose nitrogen implanted silicon have been studied for different implant and oxidation conditions in contrast to earlier work [10]. The retardation of the oxidation rate is directly related to the pile up of nitrogen to the silicon surface. The presence of a silicon oxide layer on the wafer surface has been shown to provide an effective barrier to the evaporation of nitrogen from the silicon during high temperature anneals. During the dry O_2 oxidation of low dose N_2^+ implanted silicon, the total dose of nitrogen in the silicon is involved in the inhibition of silicon oxidation. The fact that there is little or no diffusion of nitrogen through an as-grown oxide and the fact that the source of nitrogen atoms is from the silicon bulk, proves that the oxidation retardation is a surface reaction and not a diffusion limited phenomenon. Since only a reaction limited process exists, the shape of the oxidation characteristics reported here is essentially independent of the oxidation conditions such as partial pressure and temperature. For the range of nitrogen implant conditions studied, there is adequate margin to exploit the oxidation retardation properties of nitrogen implanted silicon for the use in dual thickness gate oxide applications.

CHAPTER 7

Electrical Results of CMOS Transistors with Reoxidised N_2^+ Implanted Silicon

7.1 Description of the DEC 0.5μm CMOS Process

The Digital Equipment Corporation CMOS5 process developed for the manufacture of the 300MHz 64-bit Alpha proprietary microprocessor and supporting devices uses 0.5μm minimum dimension complementary MOS transistors. The main features of this process are shallow trench isolation (STI), surface channel PMOS and NMOS transistors and 4 levels of metal wiring. A local interconnect is also used to reduce the size of the 6 transistor SRAM cell for on-chip cache area reduction. A ROXNOX gate oxide and LATID profile NMOS are used for hot carrier reliability and reduced boron penetration. A cross section of the final device is shown in schematic in Figure 7.1.

7.1.1 Shallow Trench Isolation and Transistor Formation

The starting material for the CMOS5 process is P-type epitaxial silicon from which a thin layer of silicon oxide is grown and silicon nitride deposited. The silicon nitride is patterned as are all layers in CMOS5 using i-line lithography and this forms the hard mask for the silicon etch to form the shallow trench. After a layer of silicon oxide is grown to line the trench walls, boron is implanted into the trenches to increase the parasitic MOS field transistor threshold voltage, thereby reducing leakage currents and the likelihood of latch-up. AP-CVD TEOS is deposited to fill the trench and then densified. A resist etchback process is used to semi-planarise the surface of the wafer. Chemical Mechanical Polishing (CMP) is utilised to planarise the topography further with the silicon nitride acting as an etch stop. An HF etch then recesses the TEOS in the trenches to slightly below the silicon surface and the silicon nitride is subsequently removed.

A retrograde N-well for the PMOS device is formed by the high energy implantation of P^{+++}. A sacrificial gate oxide is grown, which serves to reduce the silicon crystal surface damage from the following NMOS and PMOS threshold adjust implants and NMOS punchthrough implant. The sacrificial oxide also improves the reliability of the gate oxide at the active area to field oxide interface. After the sacrificial oxide is removed, the gate oxide is grown by a ROXNOX method to an electrically measured thickness of 90Å. Undoped polysilicon is deposited, patterned and etched. The NMOS moderately doped drain uses phosphorous LATID's to reduce asymmetry and improve hot carrier effects. The PMOS

MDD is formed using a BF_2 implant. The oxide spacers for both devices are formed by the deposition of APCVD oxide and then RIE. Self aligned source and drain implants are done next for both types of transistor. The effective channel lengths for both PMOS and NMOS are ~0.4μm. Inorder to lower the resistance between the first layer of metal and the doped polysilicon and source/drain junctions, a self aligned silicide (SALICIDE) is used. First, cobalt is sputtered onto the wafers and cobalt monocide is formed by sintering. Then the unreacted cobalt over the spacer oxide and field oxide is removed. Cobalt silicide ($CoSi_2$) is formed by a higher temperature rapid thermal anneal. A local interconnect layer was employed in the six transistor SRAM cells by the patterning of reactively sputtered TiN.

7.1.2 Interconnect Dielectrics and Metalisations.

A four level metalisation scheme is used to increase the transistor packing density in the CMOS5 process. The polycide to metal1 dielectric is deposited using a combination of Ozone TEOS, PECVD TEOS and a sacrificial boron oxide layer, to achieve the required gapfill and planarisation requirements. After the vias are etched to the cobalt silicide, a tungsten plug process is used to fill the vias due to the high aspect ratio of contact holes. The blanket tungsten process used the silane reduction of tungsten hexafluorine to deposit tungsten on a TiN seed layer. The tungsten is then etched back to stop on the oxide. Good dielectric planarisation is required so that metal stringers are avoided with as little W plug recess as possible. A TiN underlayer is sputtered for electromigration requirements and the Al 1% Cu bulk metal layer is sputtered next, followed by a thin TiN anti-reflective coating to reduce reflective notching. The metal stack is etched to define the first layer of metal wiring. The subsequent dielectric layer use a partial etchback spin-on glass scheme to compromise the trade-off between gapfill and planarity with manufacturability and throughput. The remaining metal wiring and dielectrics are similar to the ones just described up to the fourth layer of metal wiring. A thick passivation layer is deposited and etched at the bond pads. Finally, a forming gas anneal completes the fabrication process.

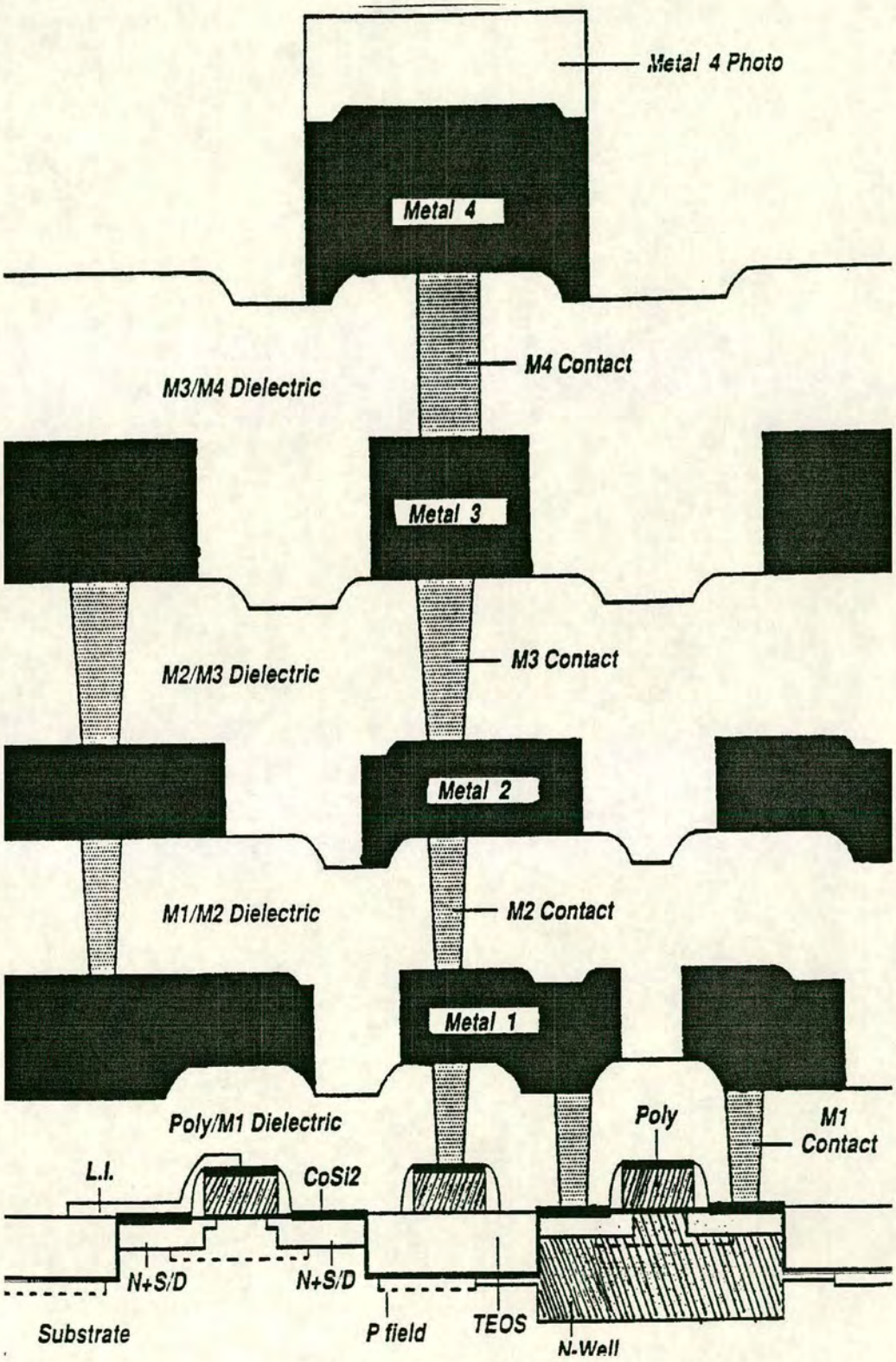


Figure 7.1 Cross-sectional representation of the CMOS5 process [166].

7.2 Experimental Details of N_2^+ Implanted Silicon as Applied to a CMOS5 Lot

A batch of 24 six inch wafers processed using the CMOS5 flow were split in order to study the effects of molecular nitrogen implantation on CMOS device performance and reliability. The mask set used was defect density chip TM50 which contained various transistor sizes, a scribe lane process monitor, antenna structures and capacitors. The nominal transistor size is 12.5μm width by 0.5μm length as drawn. The wafers were processed using the standard flow up to the point at which the threshold implants were done and the resist mask was stripped. The impact of nitrogen implantation on the field oxide areas was considered negligible. The molecular nitrogen implants were then done using the Eaton 6200 Implanter as described in Section 6.2. The implant conditions were selected based upon the fact that the dose of nitrogen in the silicon determines the oxidation retardation and that the desired energy is such that the peak of the implant is through the screen oxide but close to the silicon surface. Since the screen oxide thickness which is also the sacrificial gate oxide used in CMOS5 was 150Å, the N_2^+ Implants were adjusted for energy because the screen oxide thickness used to study the oxidation characteristics in Section 6.3 was 225Å. The energy correction was done by overlapping Figures 6.3 and 6.4 and correcting for the oxide thickness difference so that a particular oxidation retardation could be achieved. Five different molecular nitrogen implant conditions were chosen so that a wide enough range of oxidation rates could be studied. A control set of wafers (split A) with conventional oxide was also included resulting in a six way split with four wafers per split. The screen oxide was stripped using a standard 100:1 HF etch. The etch rate of the screen oxide is relatively high since it is damaged by implantation and there is unlikely to be any residual oxide.

In order to allow a comparative analysis of all six splits, the gate oxidation for each condition was altered such that an oxide of 100±5Å was grown at 900°C in dry O_2. This was justified by the fact that considering the range of oxidation rates, comparing different oxide thicknesses for a fixed oxidation time, would give different degrees of short channel effects and introduce direct tunnelling conduction for a fixed power supply voltage.

An additional constraint was placed upon the selection of molecular nitrogen implant conditions. This constraint was that the gate oxidation time should be short enough such that the threshold implants are not driven to cause a significant difference in threshold voltage. The maximum oxidation time selected was 90 minutes. Table 7.1 lists each of the six splits used in this experiment. The oxide thickness was measured using standard ellipsometry on three monitor wafers per split which were also implanted with the corresponding nitrogen

implants and oxidised in the same furnace run as the CMOS5 wafer splits. The actual oxide thickness measurement in Table 7.1 is the average of a 9-point pattern for the three monitor wafers. The oxide thickness for all splits apart from split D (calculation error) is as expected, demonstrating the predictability of this technique for different implant conditions.

The N_2^+ conditions detailed in Table 7.1 cover the range of oxidation rates in Table 6.2 from the knee of the curve, down to a point midway from the knee to the point of no oxide growth. The alphabetical ordering of the split lettering and the oxidation time show that the oxidation rate is mainly governed by the molecular nitrogen dose.

All of the wafers were deposited with polysilicon and processed using the standard CMOS5 flow up to the last high temperature drive to form the boron source and drain region. At this point one wafer from each of the six splits listed in Table 7.1 was subjected to a higher anneal temperature of 975°C for 30 minutes to study the degree of boron penetration in the PMOS device (See Section 5.4.5). The remaining five wafers per split received the normal 900°C for 30 minutes anneal which results in a very low level of boron penetration into the channel.

The lot of 24 wafers was then processed through silicidation and on up to metal1 patterning. The wafers were then annealed in forming gas to complete the fabrication process for this lot. Standard electrical testing of the scribe lane monitor showed that apart from slight variation in threshold voltage and related parametric shifts from the different thermal cycles during gate oxide growth, all wafers behaved as normal to the CMOS5 process. It is also worth noting at this point that some data is included from different lots with the ROXNOX gate dielectric. The reason for the inclusion of this data will be evident in section 7.3.5..

Split	N_2^+ Dose (cm^{-2})	N_2^+ Energy (KeV)	Oxidation Time (minutes)	Oxide Thickness (Å)
A	None	None	12	96
B	1e14	10	14	97
C	1.2e14	10	18	102
D	2e14	17	28	86
E	3e14	25	63	96
F	4e14	17	79	100

Table 7.1 Experimental Details of CMOS5 wafers implanted with N_2^+ and subsequently oxidised at 900°C in dry O_2.

7.3 MOS Device Characteristics

7.3.1 Long and Short Channel Threshold Voltage Variation

The threshold voltages V_T were measured on each wafer for 15 sites per wafer. All device measurements were done using a HP4145B Semiconductor Parameter Analyser. For the NMOS devices, the drain to source voltage V_{DS} was held at 0.05V while the gate to source voltage V_{GS} was swept from 0V to 3.3V. In all cases V_{BS}=V_S=0V. V_T was determined to be the intercept of the I_D verses V_{GS} at the point of maximum transconductance G_m. Similar measurements were made on PMOS devices with the appropriate voltage polarity.

Table 7.2 shows the average V_T for the 12.5μm square PMOS and NMOS transistor size for three wafers per split neglecting the wafers for the boron penetration study. The variations in threshold voltages within each wafer for these long channel devices were ±15 mV for all wafers. The long channel NMOS devices show an inconsistent shift in V_T with increased thermal cycle. Since the V_T was known to be very sensitive to the boron concentration at the silicon surface due to the shallow threshold adjust implant, this was surprising. The thinner gate oxide split D does however show how sensitive the V_T of both devices are to the oxide thickness. The variation of V_T due to the competing effects of increased thermal cycle and gate oxide thickness explains the inconsistent results for the long channel NMOS device. The PMOS long channel device shows a consistent V_T increase with thermal cycle for a given gate oxide thickness. This behaviour is attributed to a increase in the phosphorous concentration at the silicon surface since the peak of the PMOS threshold adjust implant is relatively deep compared to the NMOS implant.

The effective channel lengths of the nominal (12.5μm x 0.5μm) devices were measured by extrapolating the transconductance of a series of different channel lengths to find the difference between the drawn and electrical channel length. Table 7.3 shows the measured results of threshold voltage and the difference in drawn and electrical channel length (ΔL). The ΔL for both types of devices show a reduction in effective channel length with increased thermal cycle. The threshold voltage however, does not show a reduction with a slightly shorter channel length for both types of devices in Table 7.3 due to the long channel effects dominating. The difference between the long and short channel device threshold voltages from comparing each of the splits in Table 7.2 and 7.3 shows that the PMOS device follows short channel effects. In the NMOS device however, the increased threshold voltage with channel length reduction (Tables 7.2 and 7.3) is known as the reverse short channel effect (RSCE) and has been attributed to implant damage enhanced diffusion of the source/drain

implants [167]. The role of nitrogen in this enhanced diffusion is unclear at this point. The electrical width does not account for the threshold voltage variation since the nominal and square transistors were measured to be 12.0±0.3μm for both types of devices on all wafers.

Nitrogen has been previously shown to act like a donor in silicon [168]. In this case,the threshold voltage also shows a significant dependancy on nitrogen dose. Figure 7.1a and 7.1b on page 150a, show the affect of nitrogen dose such that the NMOS V_T increases and the PMOS V_T decreases with increased nitrogen dose. This is consistent with n-type doping. The threshold voltages plotted in these figures have been normalised for gate oxide thickness. The implication for this observation is that any application of the nitrogen implanted silicon technique would require V_T optimisation to componsate for the donor behaviour of the nitrogen incorporated into the silicon..

Split	N_2^+ Dose (cm^{-2})	N_2^+ Energy (KeV)	NMOS V_T (mV)	PMOS V_T (mV)	T_{ox} (Å)
A	None	None	830	-533	96
B	1e14	10	868	-527	97
C	1.2e14	10	890	-564	102
D	2e14	17	764	-479	86
E	3e14	25	807	-573	96
F	4e14	17	817	-592	100

Table 7.2 V_T variations between the process splits for the 12.5μm x 12.5μm transistor size.

Split	N_2^+ Dose (cm^{-2})	N_2^+ Energy (KeV)	NMOS V_T (mV)	PMOS V_T (mV)	NMOS ΔL (Å)	PMOS ΔL (Å)
A	None	None	848	-484	123	450
B	1e14	10	881	-478	370	817
C	1.2e14	10	912	-513	477	713
D	2e14	17	770	-423	507	910
E	3e14	25	814	-518	887	957
F	4e14	17	841	-533	837	880

Table 7.3 V_T variations between the process splits for the 12.5μm x 0.5μm transistor size.

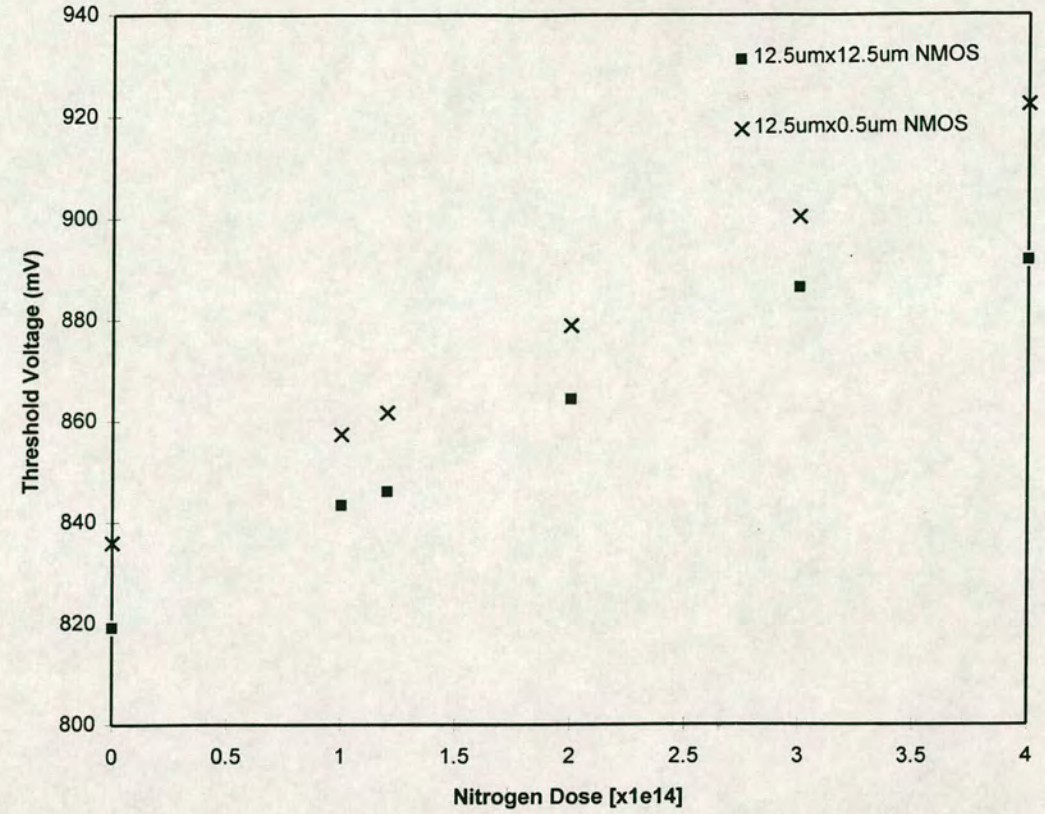


Figure 7.1a Normalised NMOS Threshold Voltage verses Nitrogen Dose.

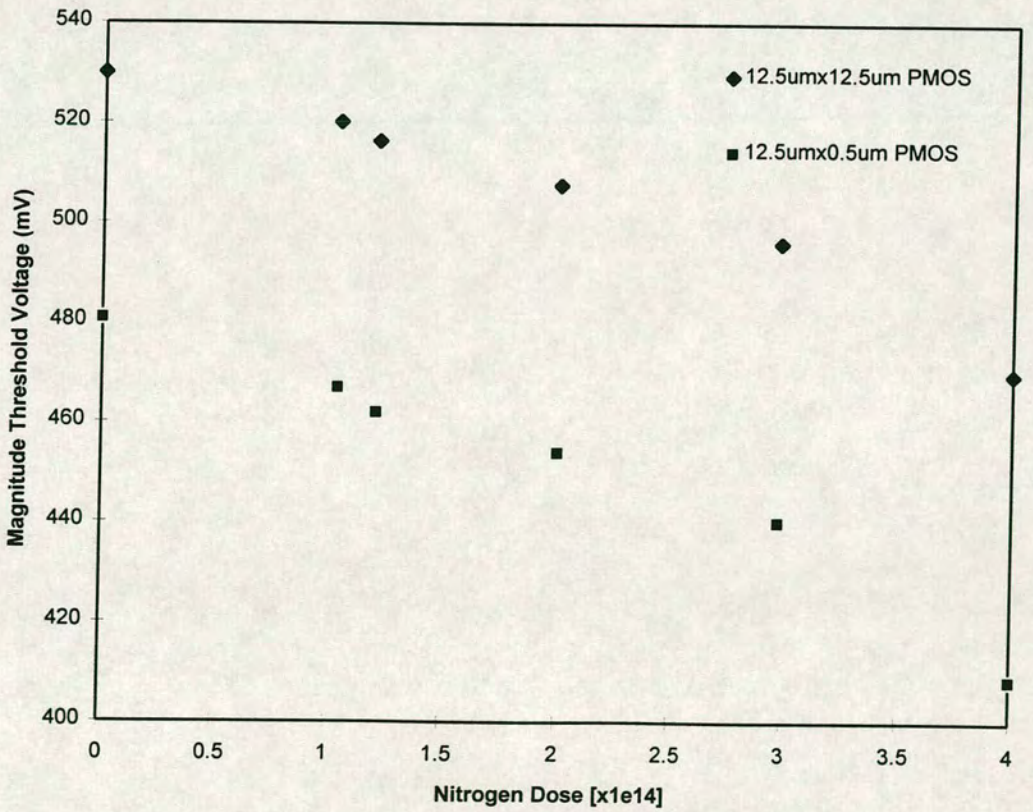


Figure 7.1b Normalised PMOS Threshold Voltage verses Nitrogen Dose.

7.3.2 Linear and Saturation Region MOS Device Characteristics

The nominal NMOS transistor current-voltage characteristics were measured by sweeping the drain voltage from 0 to 3.3V at gate voltages of 1.1V, 2.2V and 3.3V. The nominal PMOS device was measured in the same way but with the opposite voltage polarity. Figures 7.2 and 7.3 show the output characteristics for the NMOS and PMOS devices respectively. There is considerable variation in drain current at a given bias point over all the splits. Although the threshold voltage variation and effective channel length variation described in the previous section do account for some of the variation, there is still a component of drive current loss which is unaccounted for and has a larger effect with increased nitrogen content. This component of drive current loss is either a reduction in gate capacitance due to a lower dielectric constant or a reduction in channel mobility. since the percentage of nitrogen incorporated into theses gate oxides is low, the results indicate that the channel mobility is reduced due to increased nitrogen content. The nominal NMOS and PMOS device characteristics were measured for the case of drain current verses gate voltage at 0.05V and 3.3V drain voltage. Figures 7.4 and 7.5 show these plots from which the OFF state leakage and subthreshold swing can be easily determined. A low level of drain leakage current is found for all splits at V_{GS}=3.3V,V_{DS}=0.05V confirming the integrity of the gate oxide for device measurements and the low OFF current measured at V_{GS}=0.05V,V_{DS}=3.3V indicates low stand-by power consumption. The subthreshold voltage swing is determined by the reciprocal of the maximum slope of the drain current curve with gate voltage at V_{DS}=0.05V. For an ideal MOS device, the subthreshold voltage swing minimum is around 60mV/dec. The subthreshold voltage swing is a measure of how fast a device will switch from an OFF state to an ON state and it increases with the amount of interface trapped charge in the gate oxide and substrate doping. For the devices in this study, the subthreshold voltage swing is found to be 90±3 mV/dec for the NMOS and 83±2 mV/dec for the PMOS device and independent of nitrogen content. The subthreshold voltage swing for the all NMOS device splits was found to be approximately 20 mV/dec higher than that for the standard CMOS5 process with the ROXNOX gate oxide (with V_T~500mV). The PMOS subthreshold voltage swing for the nitrogen implant splits however was found to be similar to that of the ROXNOX process. The analogous behaviour of the NMOS device subthreshold voltage swing for all the nitrogen implanted silicon and conventional splits is due to the high substrate doping in the NMOS device [169]. For the purposes of this work, the fact that the conventional oxide split behaves like the molecular nitrogen implanted splits implies that the presence of nitrogen in the silicon does not deteriorate the switching speed the MOS devices.

The drain induced barrier lowering (DIBL) can be measured by the shift in subthreshold shift between the V_{DS}=0.05V and V_{DS}=3.3V curves. For both NMOS and PMOS devices, there is an equivalent shift in subthreshold slope for all splits detailed in Table 7.1. This result indicates that the short channel effects are not effected by the presence of nitrogen in these devices.

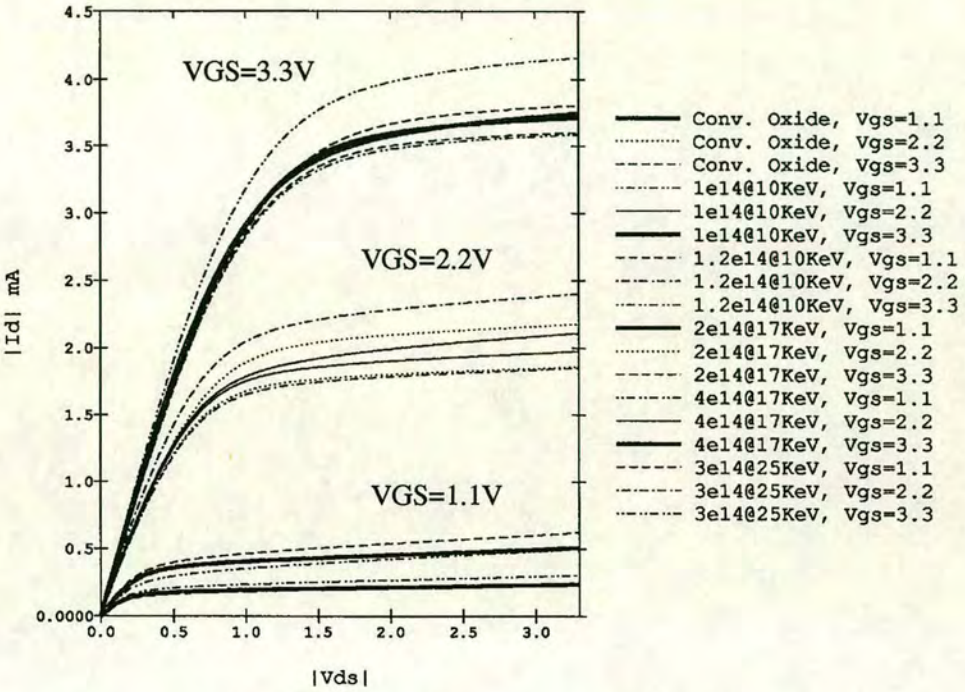


Figure 7.2 Nominal NMOS device output characteristics for each of the splits in Table 7.1.

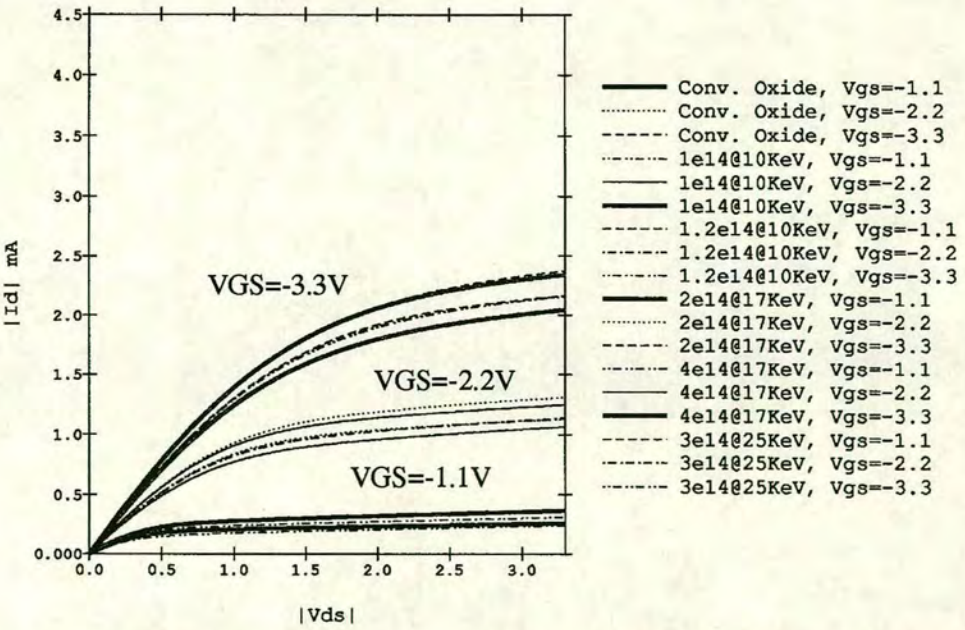


Figure 7.3 Nominal PMOS device output characteristics for each of the splits in Table 7.1.

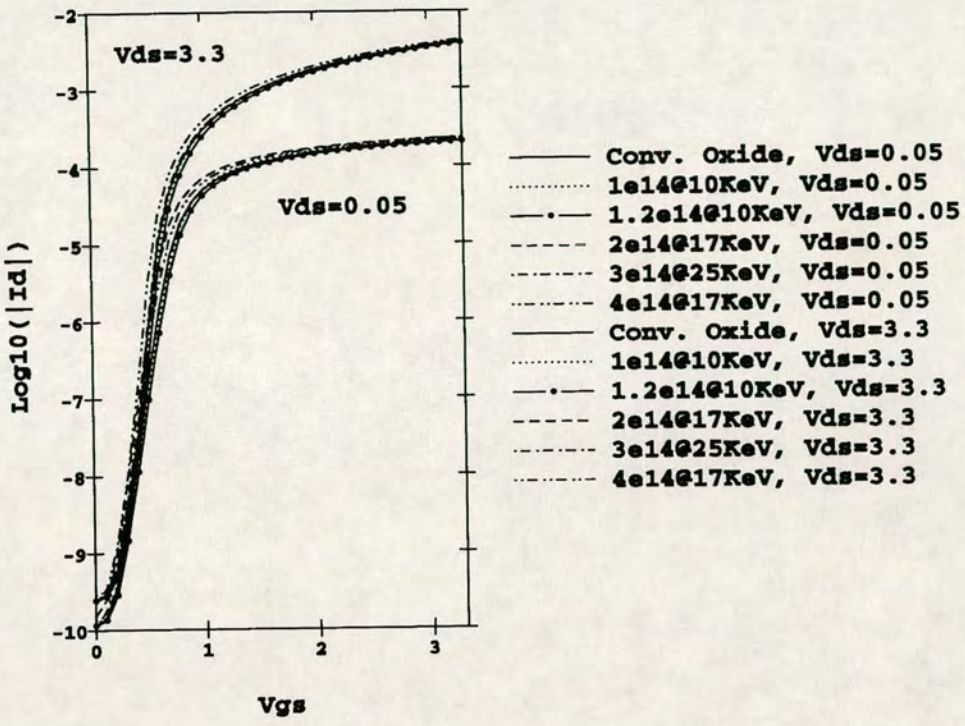


Figure 7.4 Nominal NMOS current-voltage characteristics for the splits in Table 7.1.

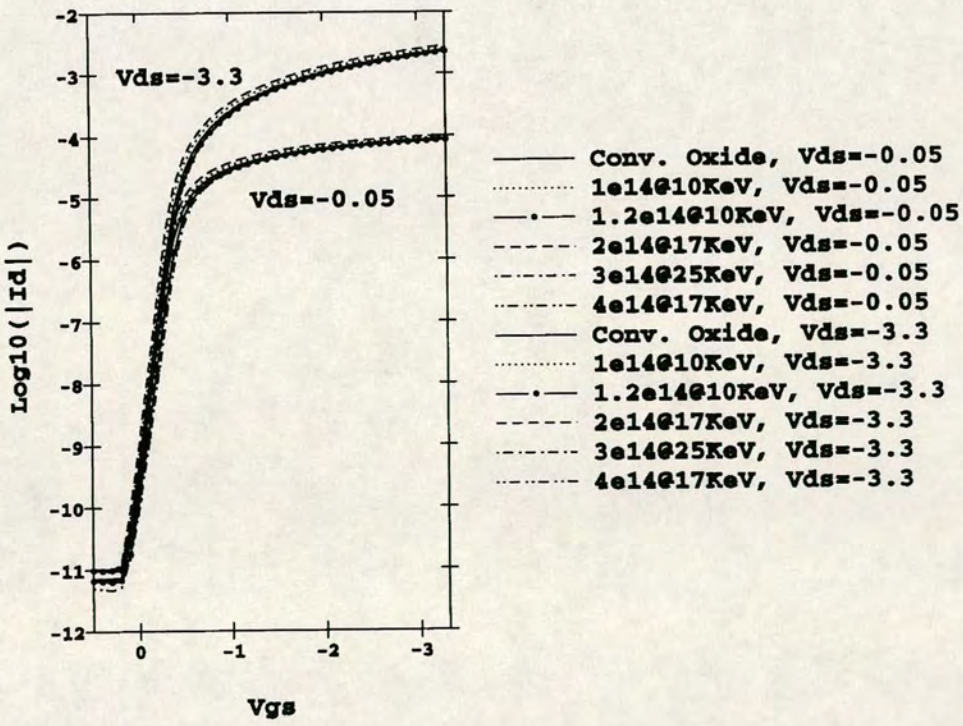


Figure 7.5 Nominal PMOS current-voltage characteristics for the splits in Table 7.1.

7.3.3 Low and High Field Mobility Characteristics

The 12.5μm square device was used to measure the mobility characteristics for both NMOS and PMOS devices as detailed in Section 3.4.3. The mobility μ was calculated from Equation 3.55 using the linear region characteristics shown in Figure 7.4 and 7.5 for the NMOS and PMOS device respectively. The effective electric field was calculated using Equation 3.74. Figures 7.6 and 7.7 show the long channel mobility verses the effective electric field for the NMOS and PMOS device respectively. Comparing the results of Figure 7.6 with Figure 3.56 shows the expected rounding of the mobility at low electric fields corresponding to the threshold voltage and the onset of the inversion layer. There is however considerable reduction of the mobility with increased nitrogen content in the silicon prior to oxidation for both the NMOS and PMOS device. This phenomenon occurs at both low and high electric fields. In other types of oxide nitridation processes, the low field mobility is know to reduce with nitridation due to an increase in the level of fixed oxide charge. These other methods however show equal or increased mobility at high electric fields compared to conventional oxides. The results in Figures 7.6 and 7.7 however show an overall reduction in mobility irrespective of the electric field but dependent on the nitrogen content in the silicon prior to gate oxidation. These results and the discussion in Section 3.4.3 imply that the mobility is reduced either because of surface scattering or lattice scattering or both. Surface scattering is due to the microroughness of the gate oxide. Lattice scattering could be due to either the presence of nitrogen in the silicon crystal or due to ion implantation damage which has not been properly annealed out.

7.3.4 Reverse Biased Diode Breakdown Results

Reverse biased diode measurements were done to determine if the cause of mobility degradation with increased nitrogen dose in the silicon was due to ion implantation damage. Large arrays (292 1μm^2 diodes) of n^+/p^--substrate and p^+/n^--well diodes were measured on 15 sites per wafer for each of the spit conditions in Table 7.1. Figures 7.8(a) and 7.8(b) show the reverse biased diode breakdown distributions. The diode breakdown was determined when there was 10nA of junction leakage. For both types of diode and for all splits, the results indicate good diode integrity which implies that any ion implantation damage has been annealed out of the active devices and is not the cause of the mobility degradation. The tighter distribution of breakdowns for the p^+/n^--well diodes is an attribute of the CMOS5 process and is related to the properties of the cobalt silicide and the source/drain junction depth.

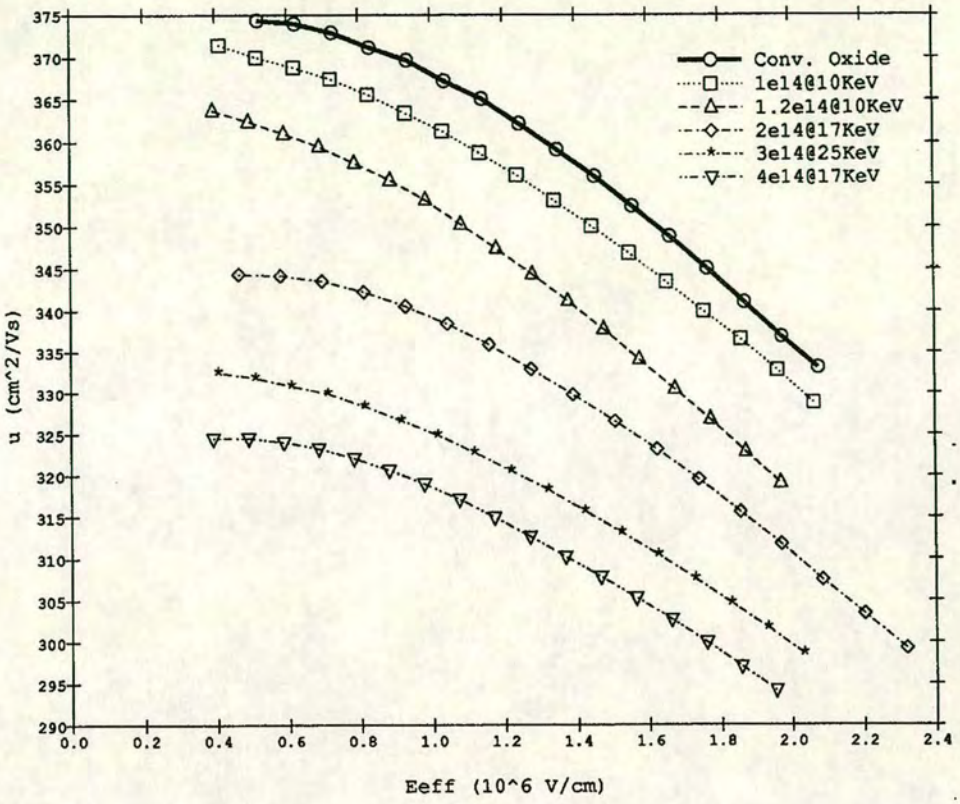


Figure 7.6 Channel mobility verses E_{eff} for 12.5µm square NMOS device.

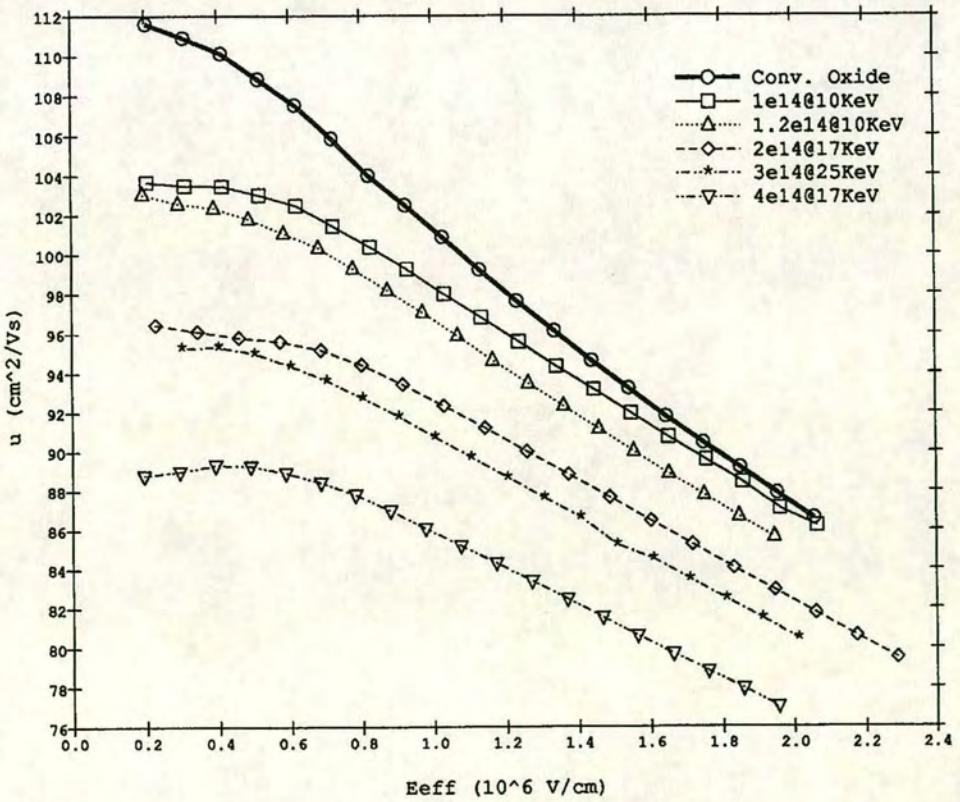


Figure 7.7 Channel mobility effective electric field for 12.5µm square PMOS device.

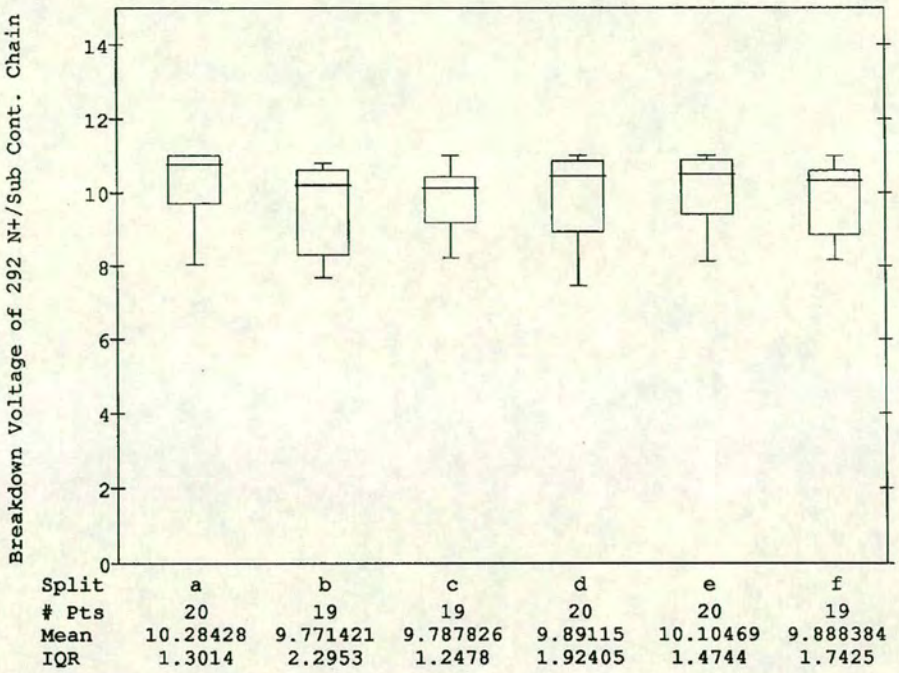


Figure 7.8(a) Reverse biased diode breakdowns for an array of 292 n^+/p^--substrate diodes for each of the splits in Table 7.1.

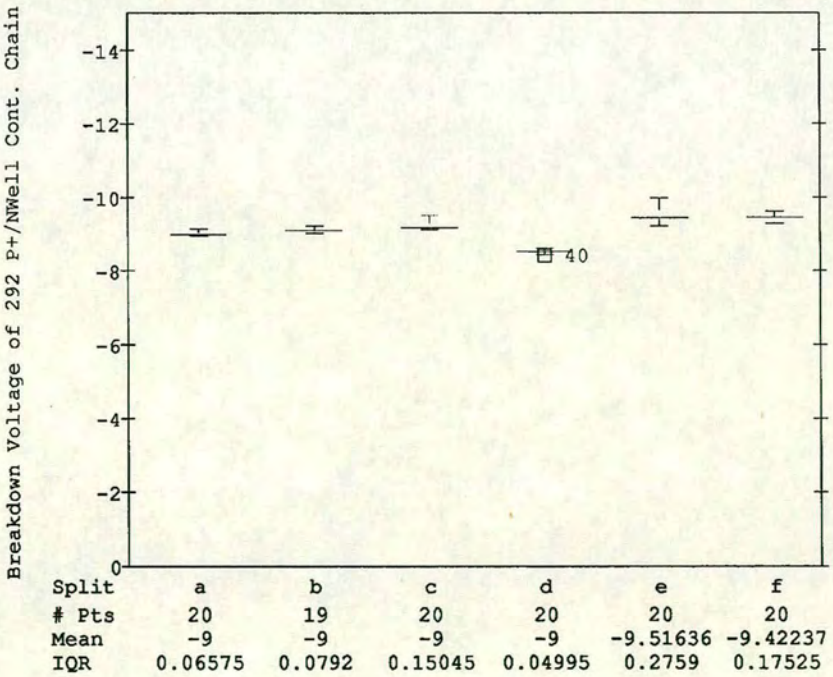


Figure 7.8(b) Reverse biased diode breakdowns for an array of 292 p^+/n^--well diodes for each of the splits in Table 7.1.

7.3.5 Results of Boron Penetration Measurements using the C-V Method

Large area capacitors were measured using the C-V technique as described in Section 3.2. The PMOS type capacitor area was 1.96e-4 cm^2. High frequency C-V measurements were taken at 1MHz using an HP4275 LCR meter. The external bias was supplied by an HP4140 picoammeter from which the displacement current could also be measured. The voltage over the capacitors was swept from inversion to accumulation with a 0.1V/sec ramp rate. A light illuminated the wafer surface just prior to the measurement, to ensure that sufficient minority carriers were present so that the device was in steady state when biased in inversion. The parasitic capacitance of the probe-station setup was subtracted from the measured data; for a C-V measurement between a gate on the front of the wafer and the back side, the parasitic capacitance was typically 350 fF. The dielectric thickness was extracted from the maximum high-frequency capacitance using a relative dielectric constant of 3.8. The electrical gate oxide thickness correlated to the measured thickness on the monitor wafers using standard ellipsometry. Figure 7.9 shows the normalised high-frequency C-V curves for the conventional oxide split wafers and each of the splits with the higher final drive condition of 975°C for 30 minutes. The amount of additional boron penetration is shown for the conventional oxide split with and without the higher temperature drive. The effect of nitrogen incorporation is clearly shown. The boron penetration is almost completely retarded for the 4e14cm^{-2} 17 KeV wafer. These results show that the reoxidised nitrogen implanted technique inhibits boron penetration with increased dose of nitrogen in the silicon.

Figure 7.10 shows the measured threshold voltage shifts between wafers with and without the higher temperature drive for both the 12.5µm square device and the nominal device. The threshold voltage shift decreases with increased nitrogen dose. The deviation in this trend for the 2e14 cm^{-2} 17 KeV wafer is due to the gate oxide thickness which is 10% thinner and hence more susceptible to boron penetration. The shorter channel length device shows a greater threshold voltage shift due to boron penetration. These results contradict a recent study [170] which showed that longer channel devices are more susceptible to boron penetration due to reverse short channel effects. The reason for this discrepancy may be due to the fact that the short channel PMOS devices in this study, do not exhibit RSCE.

The results shown here are not surprising since other methods of incorporating nitrogen into the gate dielectric such as N_2O behave in a similar manner with respect to boron penetration. As discussed in Section 5.4.4, the localised build up of nitrogen at the silicon-silicon oxide interface as measured by SIMS has been shown to be responsible for the retardation of boron penetration in these other films [146] and films studied in this work, see Section 6.2.

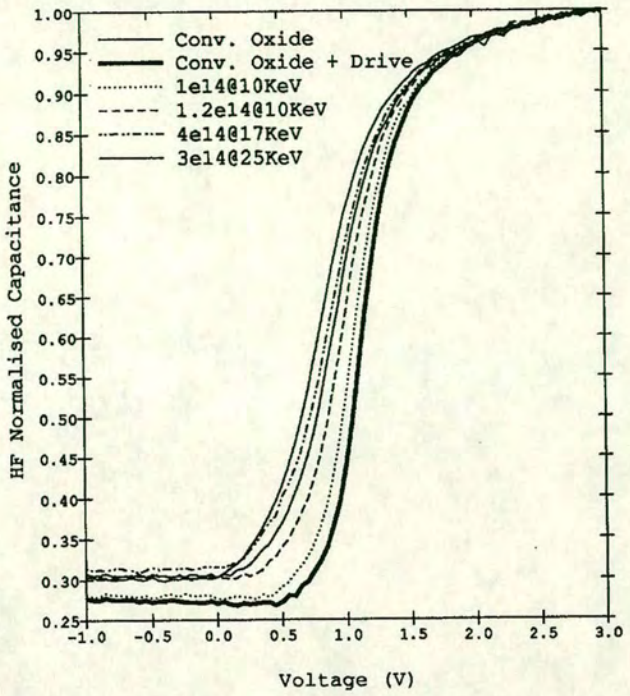


Figure 7.9 High-frequency measurements of PMOS type capacitors showing the retardation of boron penetration from flat-band shifts with increased nitrogen content.

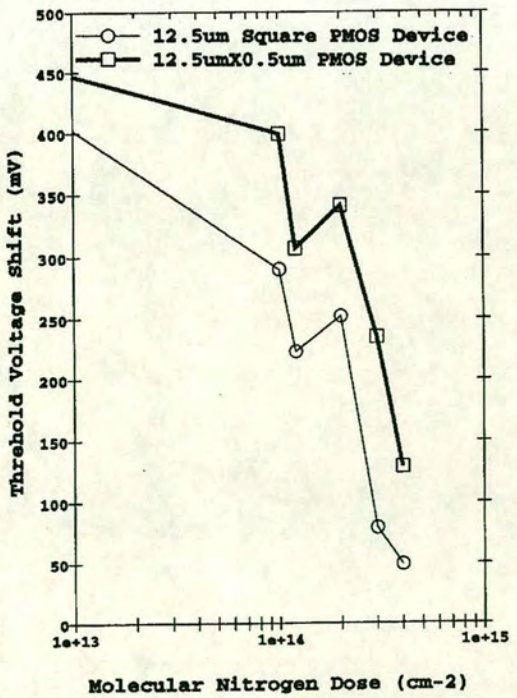


Figure 7.10 Threshold voltage shift due to boron penetration for both the 12.5µm square PMOS device (circles) and the 12.5µm by 0.5µm PMOS device (squares).

7.4 MOS Reliability Characteristics

7.4.1 Gate Oxide Breakdown Measurements

The gate oxide breakdown results were subjected to a large degree of defectivity associated with the CMOS5 process at the time this CMOS5 lot was fabricated. The defectivity level was so severe that most NMOS type large area capacitors were shorted. The large area PMOS type capacitors were defective to a less degree and so conclusions to the quality of gate oxide from the use of nitrogen implantation are relative to both conventional oxide and ROXNOX dielectric capacitors. For the case of single transistor dielectric breakdown, the comparisons are made to conventional oxide only since the yield of single transistors was high.

Oxide breakdown measurements were done using a Keithley Automatic Tester and the KLA Automatic Prober. A Time Zero Dielectric Breakdown (TZDB) test was done to determine the breakdown of the gate oxides. Since all the structures tested here were protected from plasma related degradation by diodes connected through Metal 1, the capacitors were biased into inversion for TZDB. The TZDB test consisted of 0.1 second gate voltage pulses in steps of 0.5MV/cm. Alternating low field (3.5MV/cm) measurements pulses were used to determine dielectric failure at 1μA leakage current limit.

All of the wafers in the CMOS5 lot were tested for TZDB, no correlation was found between boron penetration and gate oxide breakdown [172]. 30 sites per wafer were tested in a checkerboard pattern for each type of capacitor. Three types of large PMOS type capacitors were tested. Figure 7.11 shows the oxide breakdown for all splits and a ROXNOX representative sample of the CMOS5 process for a large array of SRAM like capacitors with an area of 3.07e-4 cm^2. There is a population of oxide breakdowns >15MV/cm which is representative of polysilicon gate depletion, this is a known attribute of the CMOS5 process. Of interest in Figure 7.11 is the reliability related oxide breakdown midfield range corresponding to 5-8 MV/cm. For the wafers with molecular nitrogen implantation there is a trend of higher midfield failures with increased nitrogen dose. The midfield failure rate increases unacceptably between 1.2e14cm^{-2} 10KeV and 2e14cm^{-2} 17KeV for this SRAM capacitor array. Figure 7.12 shows the oxide breakdown distribution for a PMOS active area/ field oxide intensive array structure with area 6.12e-4 cm^2. The midfield failure rate is slightly worse for this structure compared to the SRAM cell capacitor array and is probably due to an increase in capacitor area. Figure 7.13 shows the oxide breakdown distribution for a PMOS source/drain periphery intensive array structure with area 3.75e-4 cm^2. This

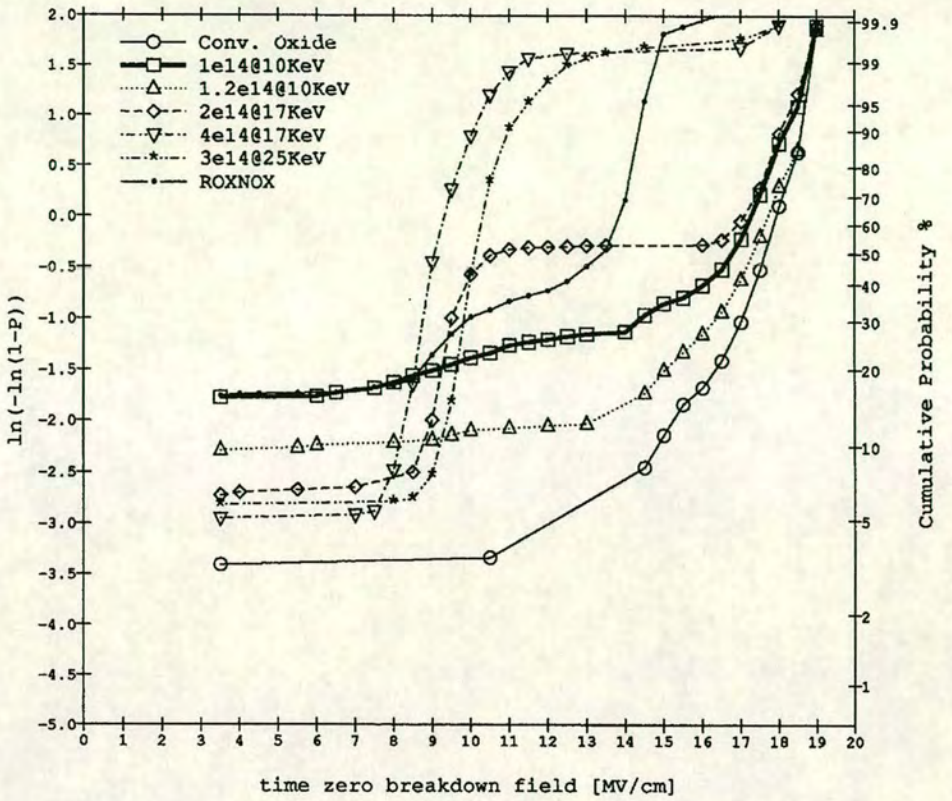


Figure 7.11 TZDB distributions for a PMOS type SRAM cell capacitor.

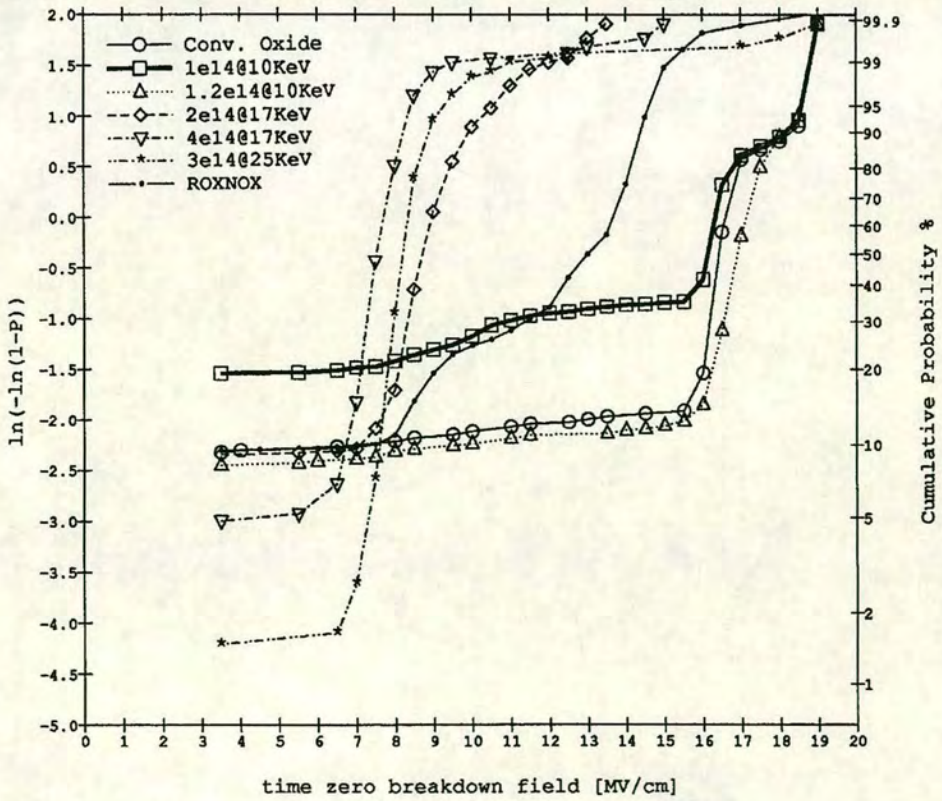


Figure 7.12 TZDB distributions for a PMOS type active/field oxide intensive area capacitor.

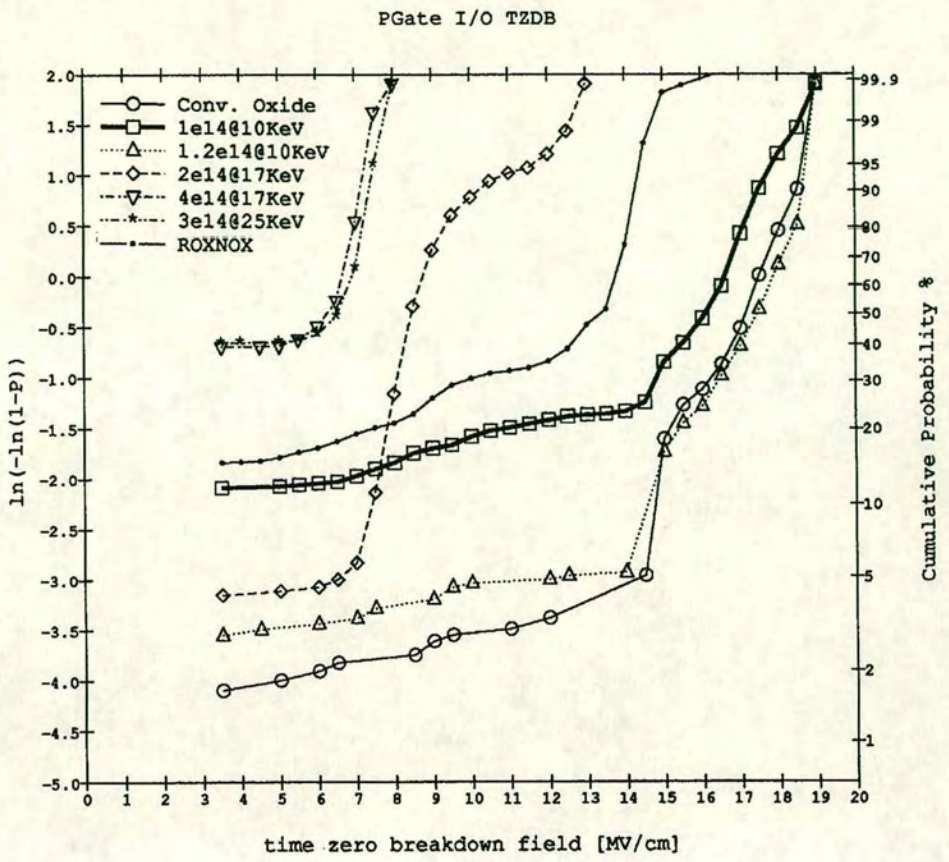


Figure 7.13 TZDB distributions of a PMOS type source/drain intensive area capacitor.

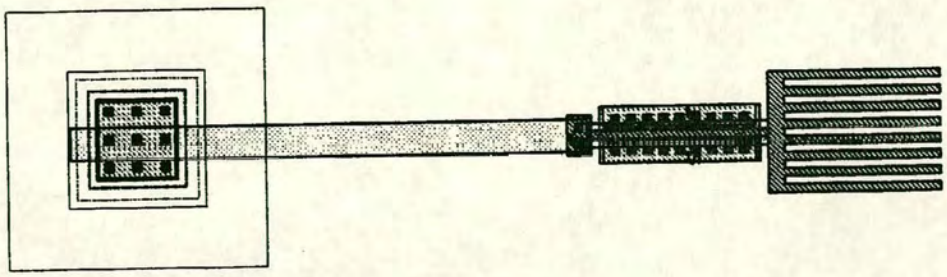


Figure 7.14 Layout of a PMOS polysilicon antenna transistor with diode protection at Metal1.

structure has a similar area to the SRAM cell and shows an even higher midfield breakdown failure rate with increased nitrogen dose in the silicon. An examination of the ROXNOX sample breakdown distribution shows that the breakdown in the midfield range is related to the active area/field oxide interface and is attributed to the gate oxide thinning around the trench corner. For the nitrogen implanted samples, the midfield failure rate is determined by the source/drain periphery and is symptomatic of plasma related charging of the gate oxide or polysilicon etching damage in the CMOS5 process as discussed in Section 4.4.

In order to determine if oxides grown from high nitrogen dose implanted silicon were more susceptible to plasma related charging, a range of single MOS transistors (10μm by 0.5μm) with various polysilicon antennas were measured for TZDB. The layout of a PMOS transistor with a polysilicon antenna and diode protection at Metal 1 is shown in Figure 7.14. NMOS and PMOS devices were measured for TZDB with polysilicon antenna ratios of 1:1, 10:1, 100:1 and 1000:1. Figures 7.15 to 7.18 show the TZDB distributions for all NMOS devices from each of the splits detailed in Table 7.1. From these figures, there can be seen to be a progressive split in the TZDB curves with increasing antenna ratio. NMOS devices with an implanted nitrogen dose higher than $2e14cm^{-2}$ exhibit lower breakdown than the devices with implanted nitrogen dose less than $1.2e14cm^{-2}$. Furthermore, the device with a higher nitrogen dose are more susceptible to plasma charging than those with a lower nitrogen dose.

TZDB distributions for PMOS devices with increasing polysilicon antenna ratio are shown in Figures 7.19 to 7.22. Again there is a trend towards a split in the TZDB curves which occurs between $1.2e14cm^{-2}$ and $2e14cm^{-2}$ and is more prominent with increased polysilicon antenna ratio. Since the effect of increasing the antenna ratio is to magnify the effects of plasma charging in the oxide, the lack of an apparent shift in the TZDB plots for the PMOS devices confirms the absence of a plasma charging problem for the PMOS device. The split in the distributions depending on the nitrogen dose is therefore indicative of an inherent difference in oxide quality which is magnified with plasma charging for the case of the NMOS device. These result are in contrast to other results which indicate that nitrogen incorporation into gate oxides reduces the susceptibility of oxides to plasma charging [171].

Although the TZDB distributions for the single NMOS and PMOS transistors shows an acceptable midfield failure rate for low polysilicon antenna ratios for all splits, the large area PMOS capacitor shows inadequate oxide reliability for all splits including the ROXNOX dielectric. The fact that the oxide reliability deteriorates with increased nitrogen dose is a major problem and puts a constraint to the application of this method for growing different thicknesses of gate oxide.

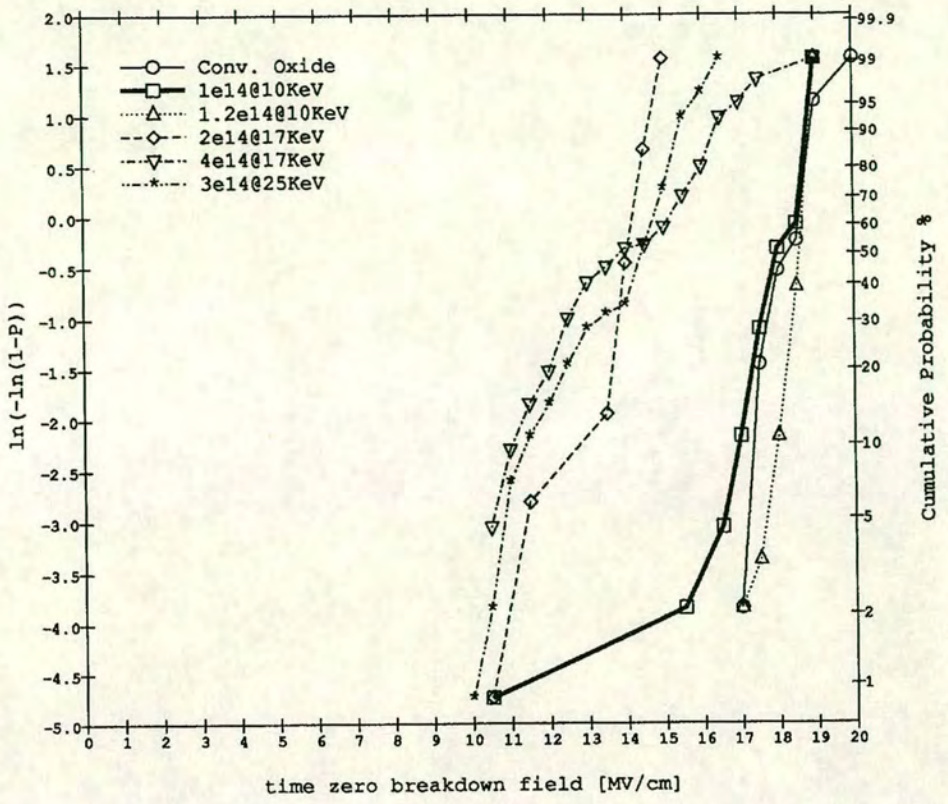


Figure 7.15 TZDB distributions of a PMOS transistor with a 1:1 polysilicon antenna ratio.

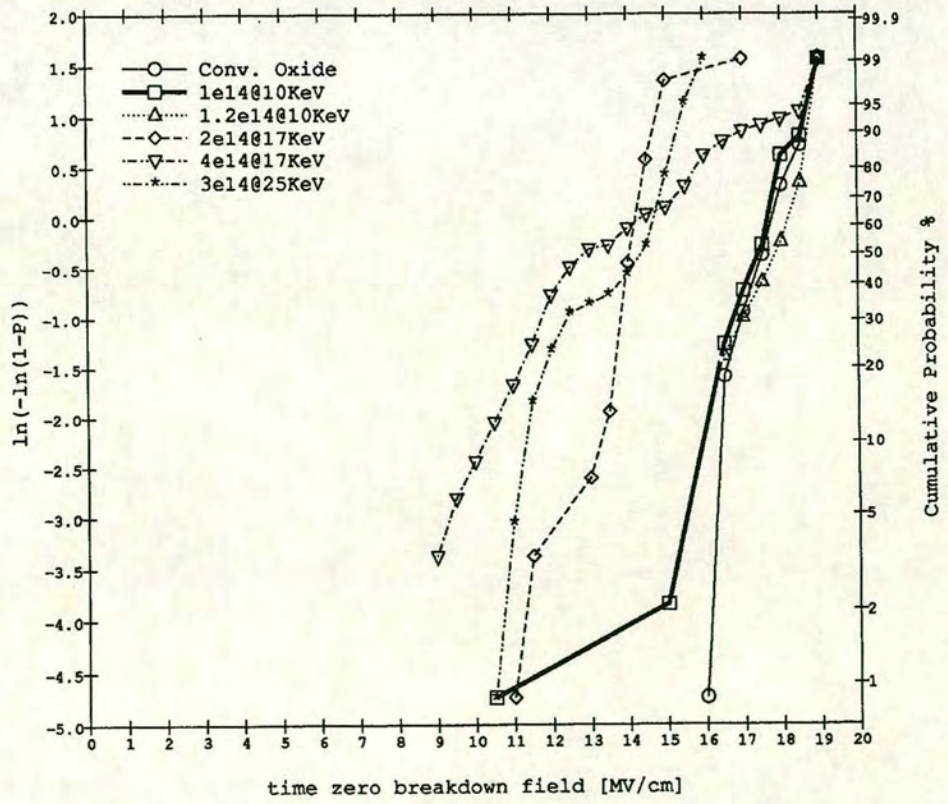


Figure 7.16 TZDB distributions of a PMOS transistor with a 1:10 polysilicon antenna ratio.

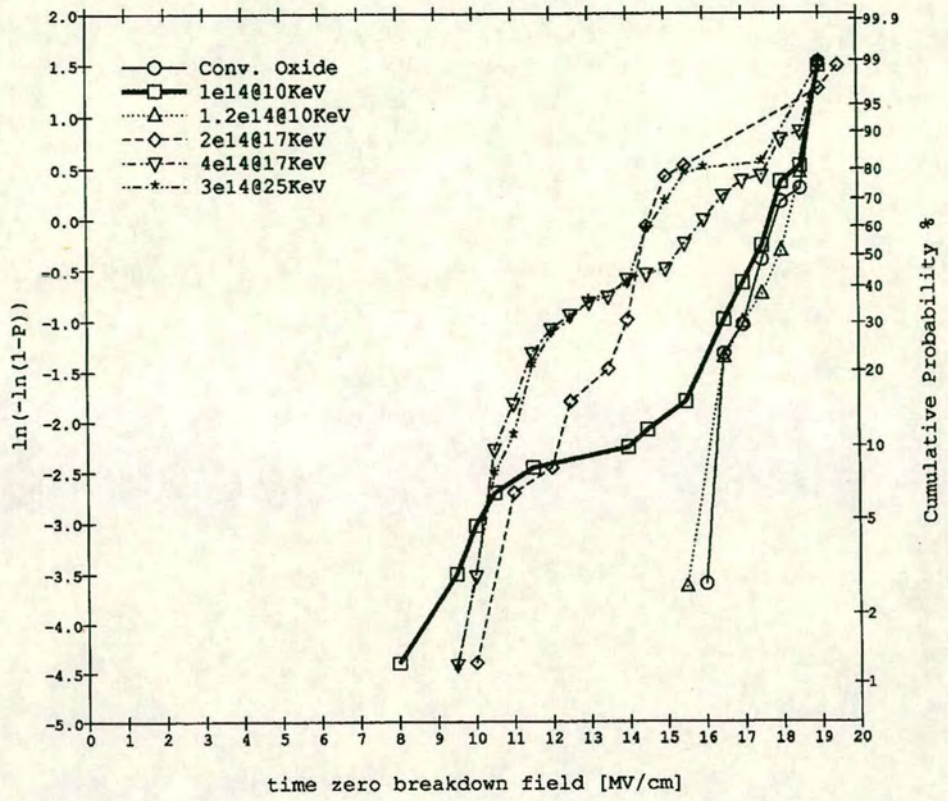


Figure 7.17 TZDB distributions of a PMOS transistor with a 1:100 polysilicon antenna ratio.

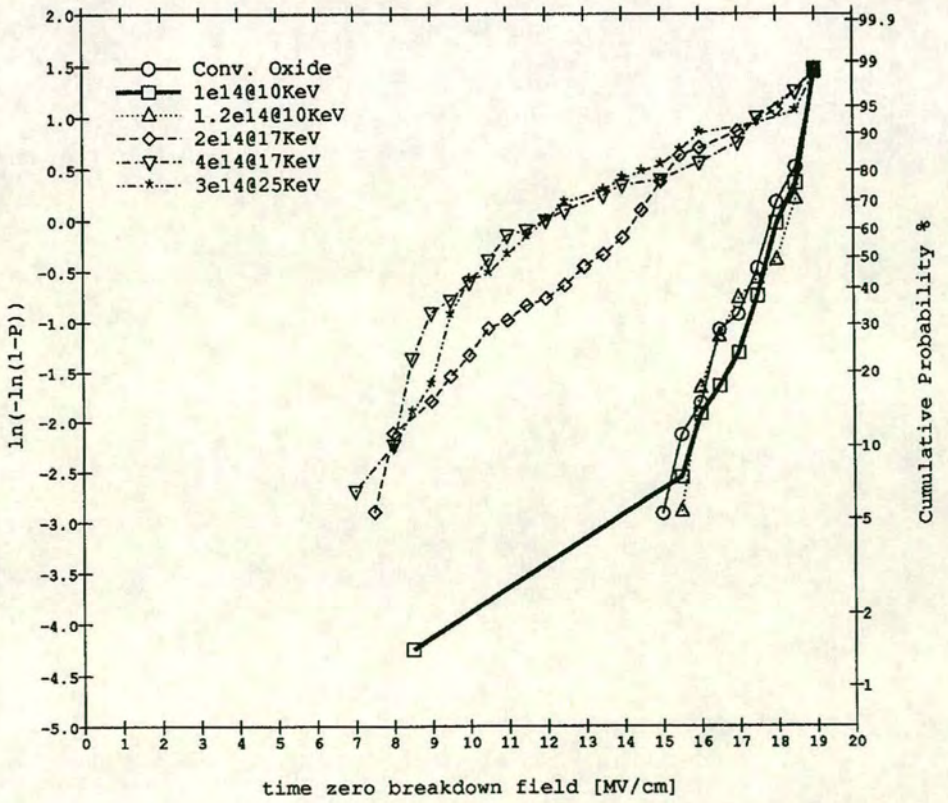


Figure 7.18 TZDB distributions of a PMOS transistor with a 1:1k polysilicon antenna ratio.

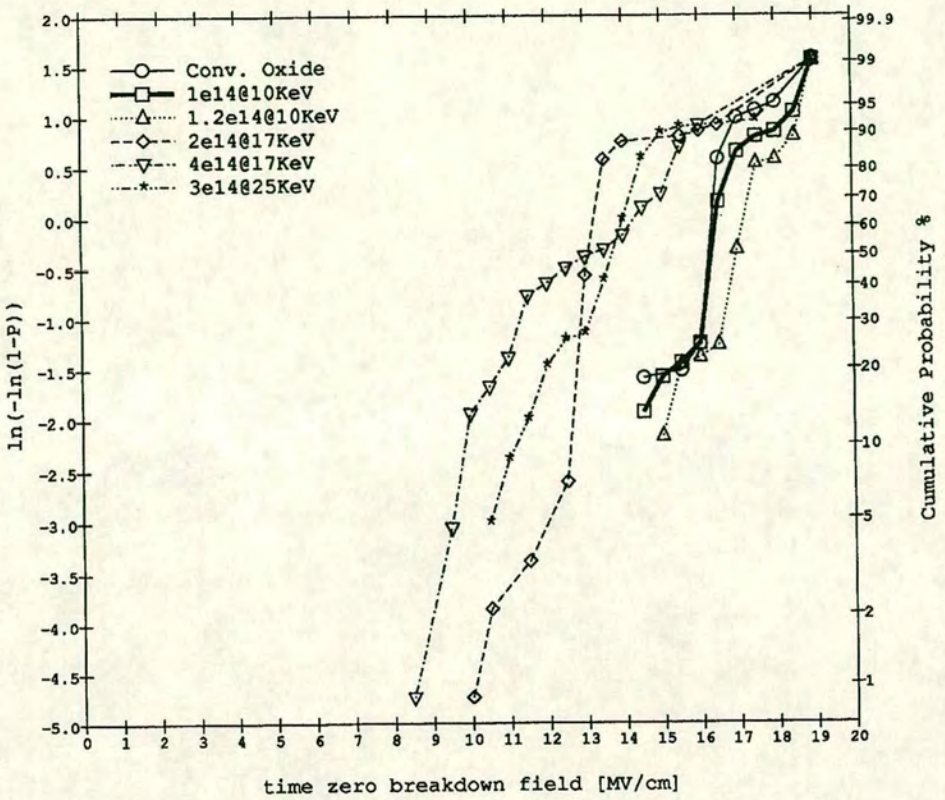


Figure 7.19 TZDB distributions of a NMOS transistor with a polysilicon antenna ratio of 1:1.

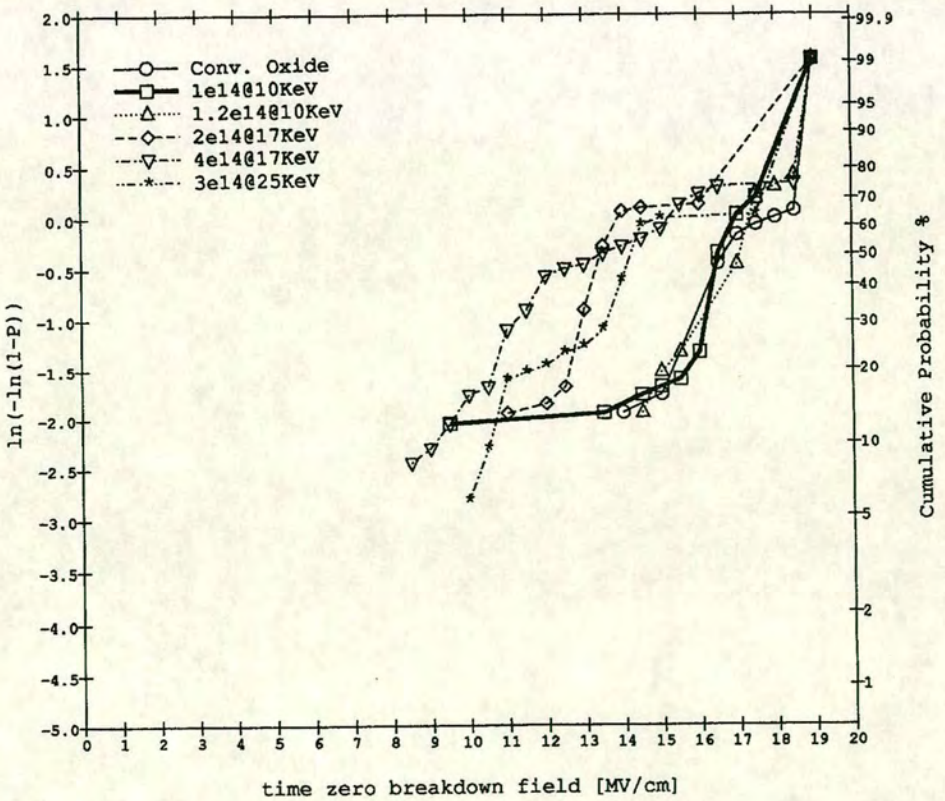


Figure 7.20 TZDB distributions of a NMOS transistor with a 1:10 polysilicon antenna ratio.

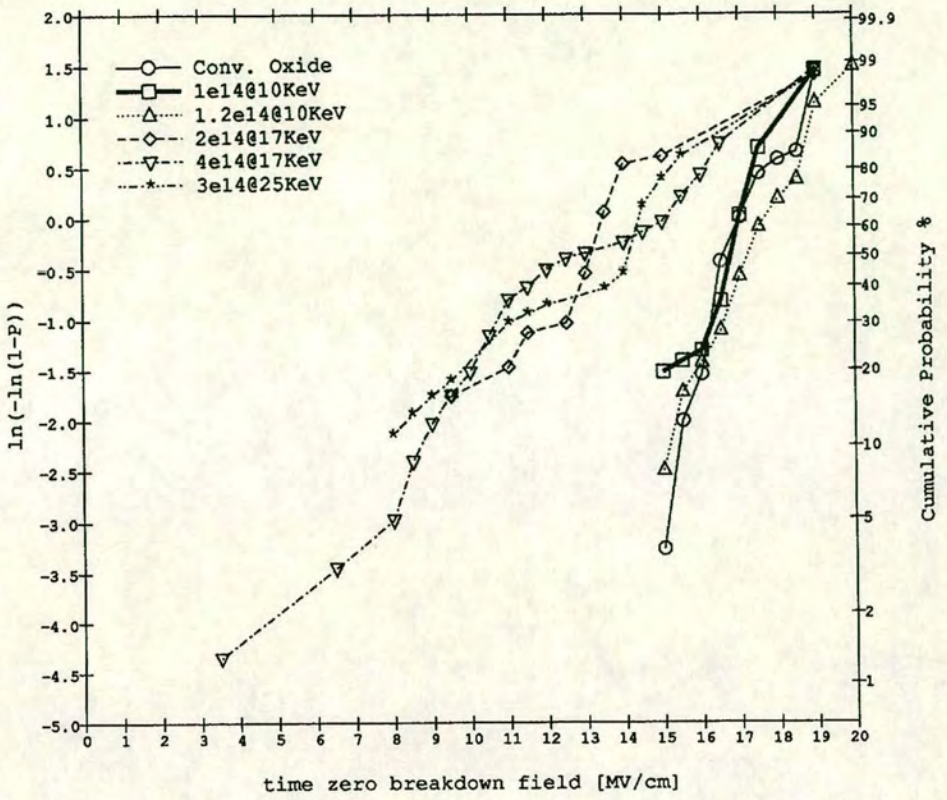


Figure 7.21 TZDB distributions of a NMOS transistor with a 1:100 polysilicon antenna ratio.

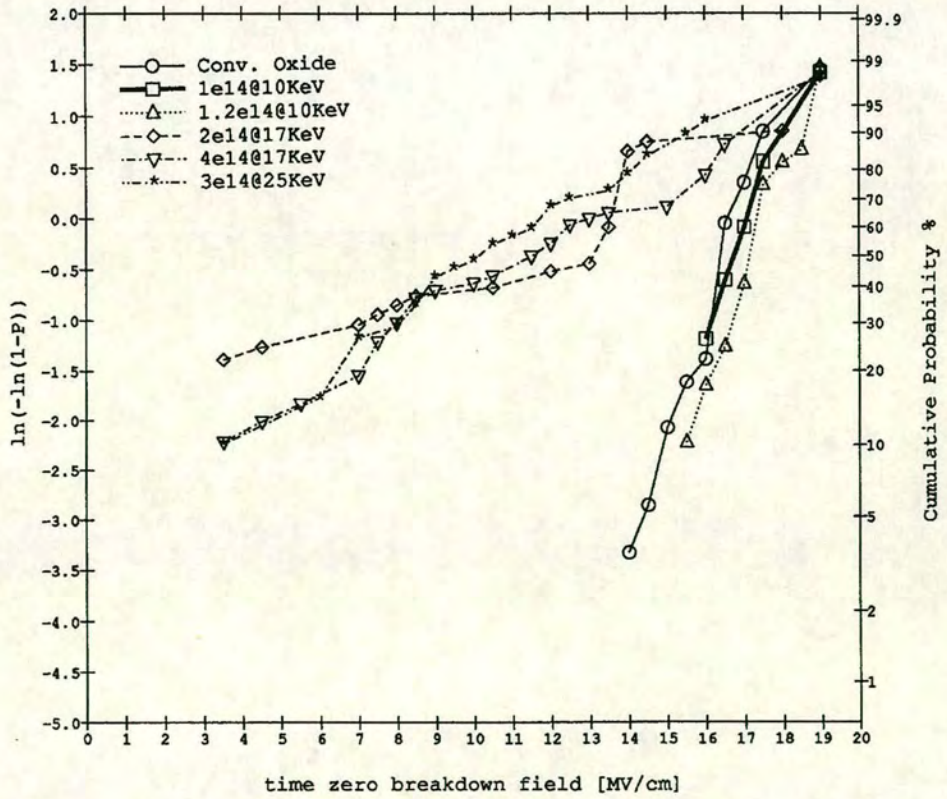


Figure 7.22 TZDB distributions of a NMOS transistor with a 1:1k polysilicon antenna ratio.

7.4.2 Hot Carrier Stress Measurements and Results

Hot carrier stress measurements were performed using an HP4145B semiconductor parameter analyser controlled by a VAX microcontroller. The VMS program used an input file to specify the stress conditions and device parameters to measure. The VMS program controlled the HP4145 so that it stressed the device under test in forward mode and then intermittently measured the device parameters of interest in both forward and reverse modes (i.e. source and drain interchanged).

NMOS and PMOS device were selected with L_{eff}=0.35μm for all splits measured. The transistors were diode protected at Metal 1 so that the effect of plasma charging damage on the hot carrier robustness would be reduced. The methodology for hot carrier stressing was as discussed in Section 5.1.5. NMOS devices were stressed in each of the three modes of stress. For the low gate voltage stress conditions associated with hole trapping in the gate oxide, the devices were stressed with V_{DS}=5.3, 5.5, 5.7 or 5.9V and $V_{GS}=V_{DS}/5$. The condition of enhanced interface trapping stress of the gate oxide was measured for V_{DS}=5.3, 5.5, 5.7 or 5.9V and V_{GS} was determined as the point of maximum substrate current. The third condition of NMOS stress was for $V_{DS}=V_{GS}$=5.3, 5.5, 5.7 or 5.9V which is associated with device parametric shifts due to electron trapping in the gate oxide. PMOS devices were stressed under conditions of maximum electron trapping in the gate oxide for V_{DS}=-5.3, -5.5, -5.7 or -5.9V and V_{GS} was determined as the bias point of maximum gate current.

Due to the TZDB results in the previous section and a time constraint, only splits A,B,C and in some cases F, were stressed for 10^6 seconds. Curve fitting extrapolations were made to estimate the accelerated lifetime of devices beyond 10^6 seconds. The parameters which were measured for hot carrier degradation were the forward and reverse saturation currents. The criterion for device failure was obtained for the forward saturation current degradation by 5% and the reverse saturation current degradation by 15% for the NMOS devices. In the case of the PMOS devices, the failure criterion was determined when the saturation currents increased by either 5% or 15% for the forward and reverse case respectively. These criterion for failure are based upon circuit simulation studies [89]. Since an extensive study of the CMOS5 hot carrier reliability characteristics had been done [172], straight lines with the appropriate gradient were fitted to the data points where it was appropriate to do so, such that the differences between the splits in terms of hot carrier robustness is interpreted by the shift in the intercept of the y axis. For the cases of data which behave differently to the characterisation study [172], a line was fitted to the data points using a least squares fit.

Figure 7.23 and 7.24 show the NMOS hot carrier data for the forward and reverse stress under the conditions of enhanced hole trapping in the gate oxide. The axes are normalised to either forward or reverse saturation current according to Equation 5.9. The effect of nitrogen incorporation during the gate oxidation process is shown by the steeper line fit which increase the sensitivity of these devices to hole trapping parametric shifts. The result of extrapolating the accelerated conditions back to worst case operating conditions of V_{DS}=4.1V corresponding to a $\log_{10}(I_b/I_d)$=-1.65 is that the nitrogen implanted splits have a longer hole trapping limited lifetime compared to the conventional oxide split. The reason for the greater sensitivity of accelerated hole trapping conditions for the nitrogen implanted silicon splits is thought to be due to a higher threshold voltage so that at higher drain stresses, there is a greater sensitivity to drain induced barrier lowering.

Figures 7.25 and 7.26 shows the accelerated stress results for the conditions of maximum interface trap generation in forward and reverse modes respectively. Although the forward mode results show only a slight dependence on nitrogen content, the reverse mode stress results show a significant improvement of interface trap related hot carrier robustness. Since the limiting hot carrier stress conditions for CMOS5 were known to be due to reverse saturation current degredation during maximum interface trap generation in NMOS devices, this result is very important. The improvement of reverse saturation current degradation for these conditions is attributed to a lower concentration of newly formed interface dangling bonds with incremental hot carrier stress, due to the substitution of nitrogen atoms for hydrogen atoms at the silicon-silicon oxide interface. The bond strength of Si-N is considerably higher than that of Si-H as discussed in Section 5.4.4. The fact that a reasonably good fit is achieved for all the splits with the same gradient as other CMOS5 hot carrier studies implies that the same mechanism for parametric shifts is occurring. The improvement in device lifetime for the 4e14cm^{-2} at 17KeV over the conventional oxide, shown by the difference in intercept of the y-axis, results in a ~2 times improvement in hot carrier lifetime.

The hot carrier lifetime under the conditions of enhanced electron trapping is shown in Figures 7.27 and 7.28 for the forward and reverse modes respectively. The axes have been normalised to the respective drive current according to Equation 5.10. The results show that there is little significant difference between the splits measured. This result is encouraging since the electron trapping lifetime of other methods of nitrogen incorporation such as N_2O nitridation show a decrease in the hot carrier lifetime under these conditions [173]. The absence of a shift in electron trapping device degredation is thought to be due to little

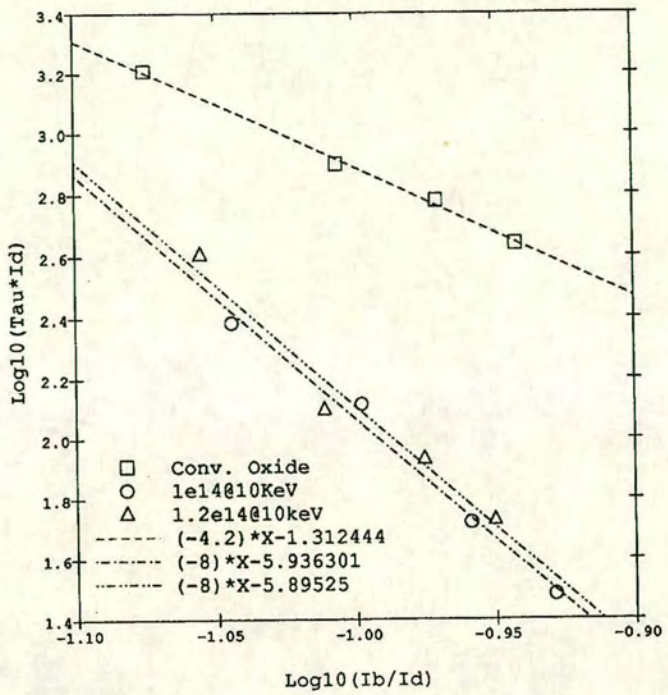


Figure 7.23 Hot carrier stress degradation of the forward saturation current for the conditions of enhanced hole trapping in the NMOS device with L_{eff}=0.35μm.

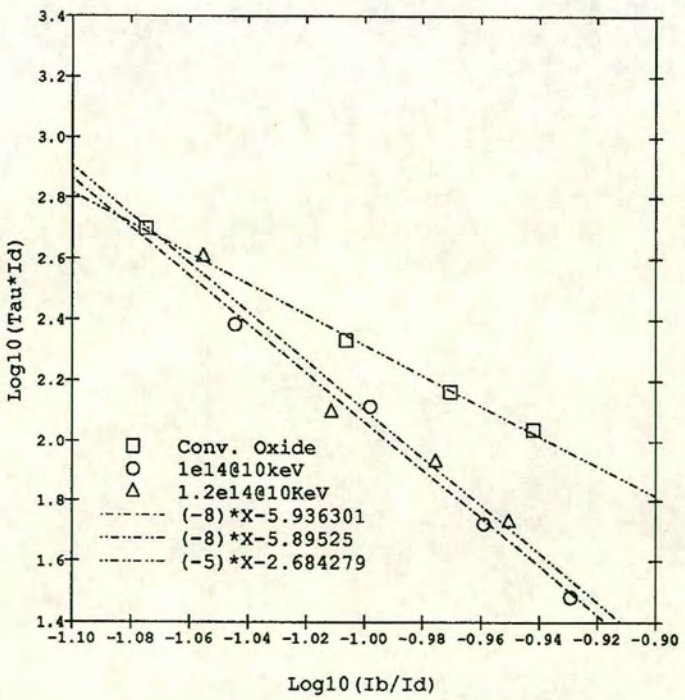


Figure 7.24 Hot carrier stress degradation of the reverse saturation current for the conditions of enhanced hole trapping in the NMOS device with L_{eff}=0.35μm.

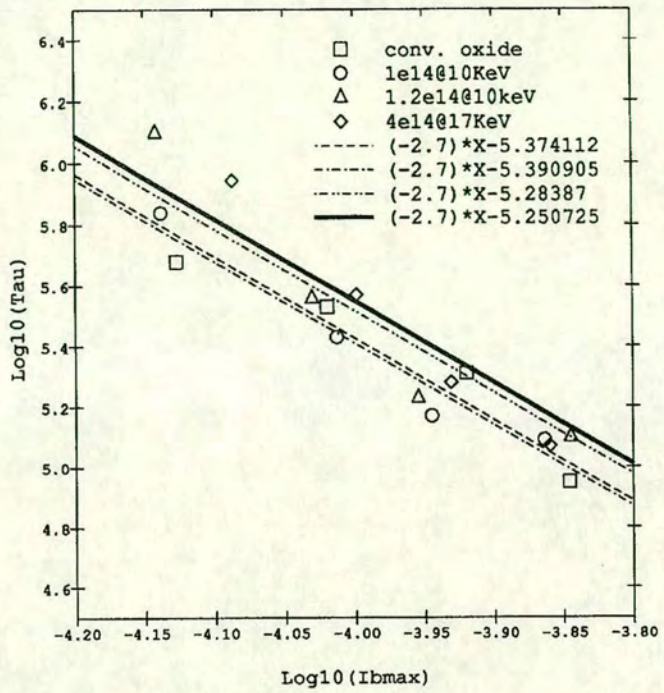


Figure 7.25 Hot carrier stress degradation of the forward saturation current for the conditions of enhanced interface trap generation in the NMOS device with L_{eff}=0.35μm.

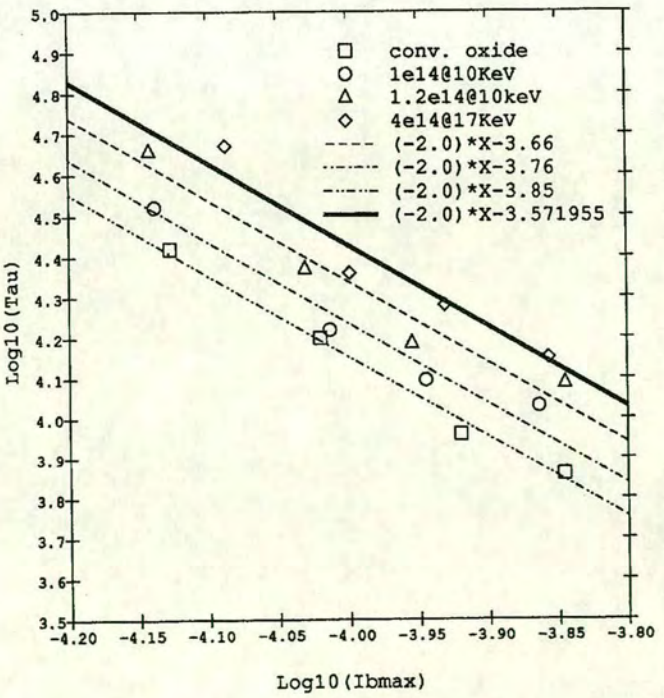


Figure 7.26 Hot carrier stress degradation of the reverse saturation current for the conditions of enhanced interface trap generation in the NMOS device with L_{eff}=0.35μm.

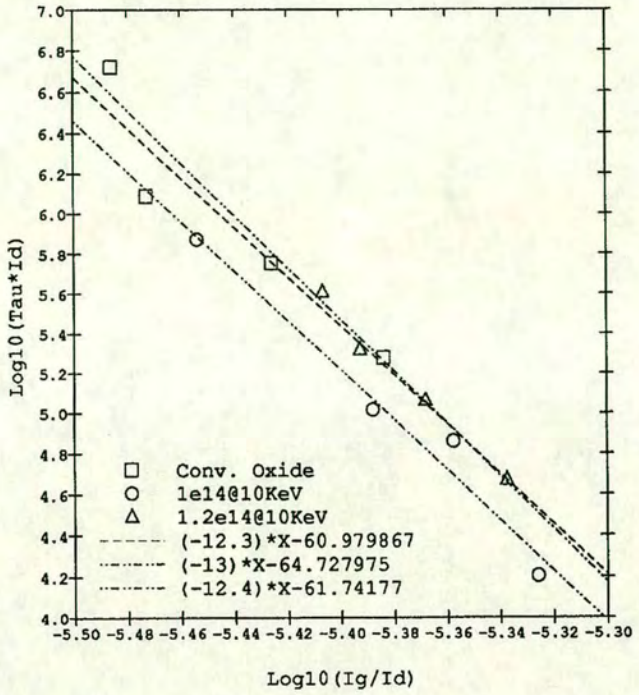


Figure 7.27 Hot carrier stress degradation of the forward saturation current for the conditions of enhanced electron trapping in the NMOS device with L_{eff}=0.35μm.

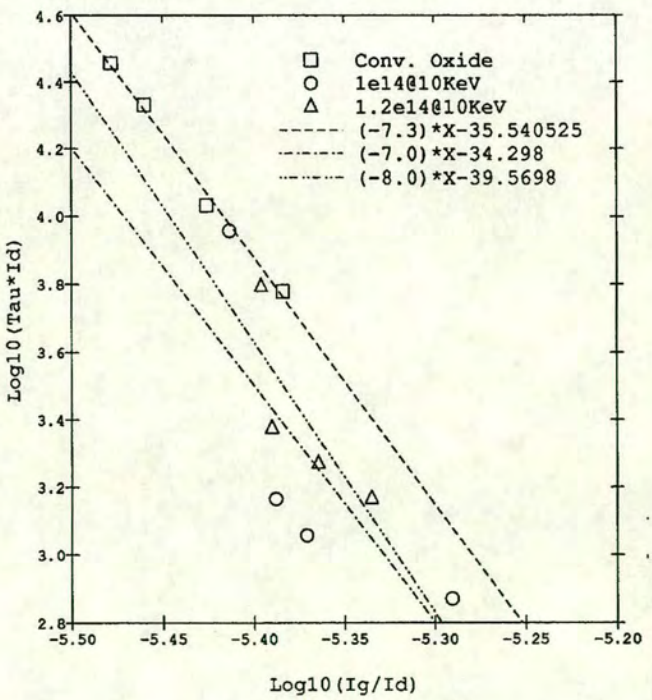


Figure 7.28 Hot carrier stress degradation of the reverse saturation current for the conditions of enhanced electron trapping in the NMOS device with L_{eff}=0.35μm.

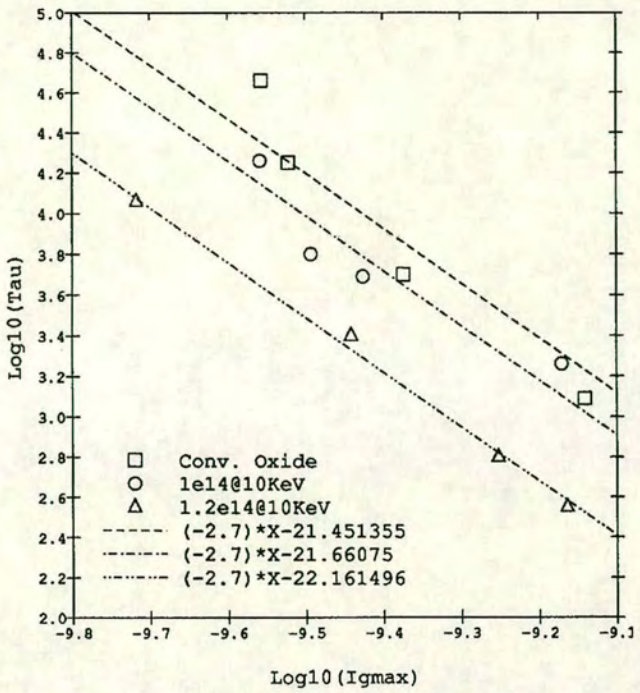


Figure 7.29 Hot carrier stress degradation of the forward saturation current for the conditions of maximum electron trapping in the PMOS device with L_{eff}=0.35μm.

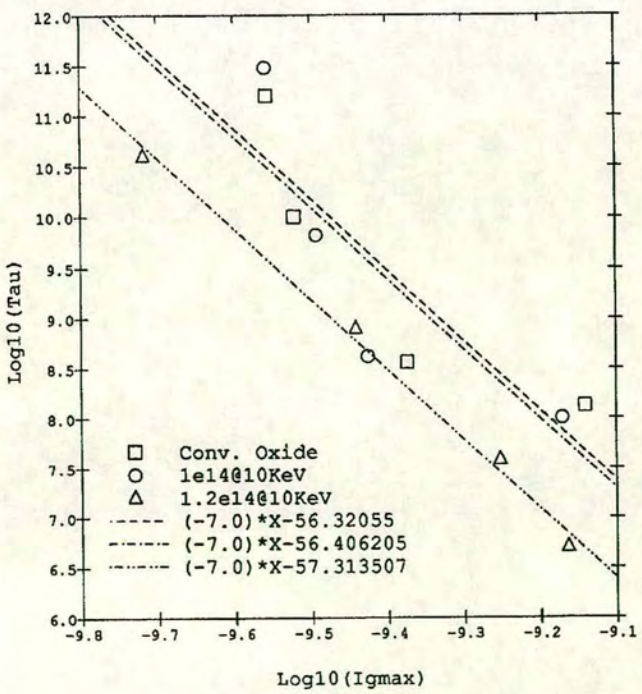


Figure 7.30 Hot carrier stress degradation of the reverse saturation current for the conditions of maximum electron trapping in the PMOS device with L_{eff}=0.35μm.

difference in fixed oxide charge between the splits.

The results of the PMOS hot carrier stressing under conditions of maximum electron trapping in the gate oxide is shown in Figures 7.29 and 7.30 for the forward and reverse modes respectively. As would be expected from the results of enhanced electron trapping in the NMOS device, there is little shift in the PMOS device saturation currents with increasing nitrogen content.

Extrapolations back to worse case operating conditions of V_{DS}=4.1V for all three modes of stress on the NMOS device and for V_{DS}=-4.1V for the single stress mode on the PMOS device, result in the limiting stress condition to achieve 10 years operating lifetime being the maximum interface trap generation condition on the NMOS device. The reverse saturation current degradation determines the point of the performance/reliability trade-off as it does for CMOS5 in general. The fact that this constraint condition has not changed from the CMOS5 process to the present work implies that the same physical mechanism for hot carrier degradation can be monitored or improved to determine the point of optimum performance and reliability. Although this hot carrier reliability study is incomplete, the effect of increasing nitrogen content in this work seems to improve the interface trap dominated hot carrier reliability in a trade-off with gate oxide reliability and drive current.

7.4.3 Results of Gate Oxide Microroughness Measurements

As mentioned in Section 7.2, a group of three wafers per split was used to measure the representative physical thickness of the oxide on the CMOS5 lot. These monitor wafers were also p-type silicon but were Czochralski silicon compared to the p-type epitaxial silicon wafers for the CMOS5 lot. The following work assumes that there is no difference between the monitor wafers and the CMOS5 p-type epitaxial silicon wafers in terms of thin oxide growth properties in the presence of molecular nitrogen implanted silicon.

Tapping mode Atomic Force Microscopy (AFM) was used to image a 1μm by 1μm area from the centre of one wafer from all six splits using a Nanoscope III SPM system [174]. Tip apex radii used was <20nm. The resolution from each wafer was 512 by 512 pixels with a scan rate of 1 line per second. Root Mean Square (RMS) roughness were obtained from these images using in-house software. The oxide surface and the substrate surface after a 1 minute dip in 10:1 HF to etch the oxide off, were profiled. Table 7.4 shows the RMS microroughness of the oxide and silicon surfaces for each split. There is a clear trend of increased microroughness with nitrogen content in the silicon for both the oxide and silicon substrate surfaces. The silicon substrate is consistently smoother than the oxide surface.

Split	N_2^+ Dose (cm^{-2})	N_2^+ Energy (KeV)	RMS of oxide surface (nm)	RMS of substrate surface (nm)
A	None	None	0.165-0.176	0.153-0.159
B	1e14	10	0.192-0.198	0.178-0.185
C	1.2e14	10	0.211-0.212	0.179-0.181
D	2e14	17	0.276-0.282	0.237-0.269
E	3e14	25	0.518-0.519	0.392-0.401
F	4e14	17	0.530-0.554	0.415-0.433

Table 7.4 RMS microroughness measurements using AFM on the oxide and silicon surfaces.

. This difference indicates that the oxide is non-uniform in thickness over a short range order. These results compare to values of 0.183-0.185 nm and 0.130-0.135 nm for previous measurements of conventional oxide and silicon surface microroughness done on p-type monitor wafers [175].

Figures 7.31 to 7.36 show the AFM topographic images of the silicon substrate surface with an amorphous polysilicon capping layer, for each of the splits in Table 7.4. There is clearly an increased silicon microroughness with increasing molecular nitrogen dose in the silicon. In order to verify that the oxide growth is non-uniform with increased molecular nitrogen dose, cross-sectional Transmission Electron Microscope (TEM) images were taken of splits A, E and F. These images are shown in Figures 7.37 to 7.39 respectively. The TEM images were obtained by tilting the samples to the [011] zone using a Philips CM30 TEM operating at 300 KV. The images were collected with a Gatan CCD camera with a resolution of 1024 by 1024. The oxide thickness was measured at different locations on the images using the Si lattice fringes as calibration. Figure 7.37 shows a uniform oxide thickness and no silicon surface pitting as one would expect from the conventional oxide split. For the relatively high molecular nitrogen doses however, Figures 7.38 and 7.39 show localised thinning of the oxide and an increased silicon microroughness. Measurements of the oxide thickness over a ~1000Å field of view indicate an oxide thickness variation of >20Å for Figure 7.38. These observations indicate that the microroughness is due to local variations in the rate of oxidation. This results in a non-uniform oxide thickness and a lack of conformality of the oxide with the silicon substrate. These results are explained by proposing that the

mechanism for non-uniform oxidation rate with increased nitrogen dose in the silicon is due to the non-uniform redistribution of nitrogen in the silicon during oxidation. This result in localised pockets of silicon which have different concentrations of nitrogen such that the oxidation rate is affected. The effect of increased nitrogen dose is to increase the standard deviation of localised nitrogen content in the silicon during oxidation.

The results of this study into the microroughness of oxides grown on nitrogen implanted silicon have other implications. This work combined with the oxide breakdown results given in Section 7.4.1 show a correlation of oxide breakdown with increased oxide microroughness. It is possible to determine the point at which gate oxide breakdown will be degraded by microroughness from pre-oxidation cleans or poor oxide growth conditions. These results have deconfounded other work [176] on the affect of pre-oxidation cleans on the silicon microroughness and dielectric breakdown due to the elimination metallic impurities from the experimental conditions. The increased silicon oxide/silicon interface microroughness also correlates to the deterioration in channel mobility as shown in Figures 7.6 and 7.7. This is intuitively correct since the scattering of channel carriers in both surface channel NMOS and PMOS devices will increase with silicon surface microroughness.

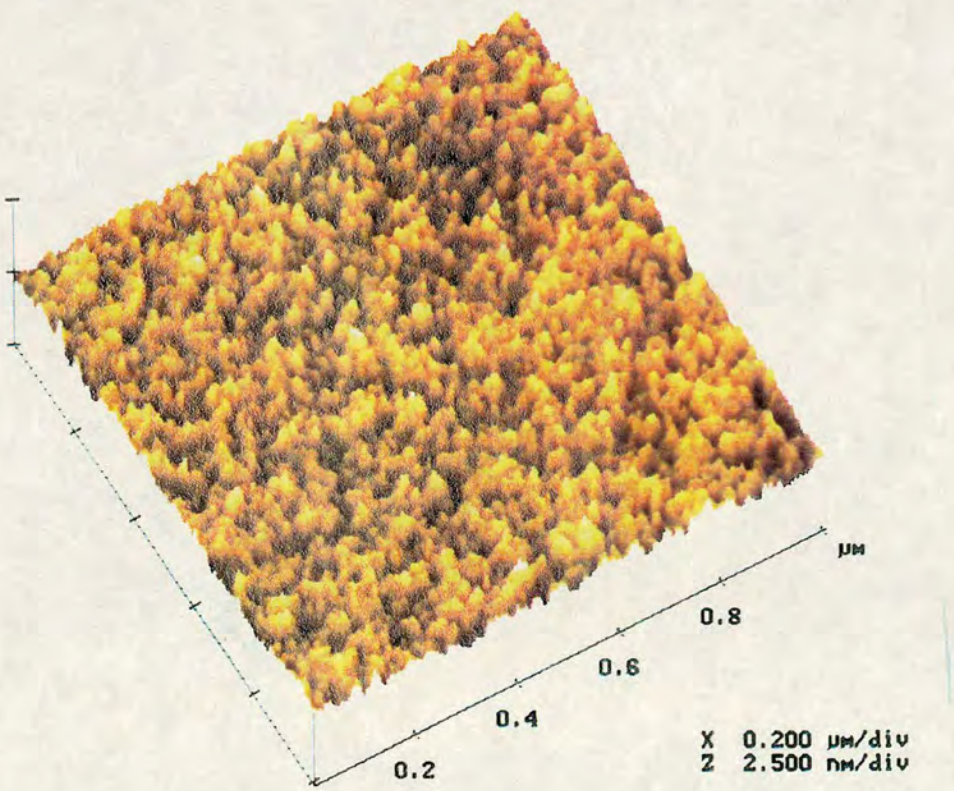


Figure 7.31 Topographic AFM image of the silicon surface with a 1μm by 1μm area and Z range of 5 nm for the conventional oxide sample.

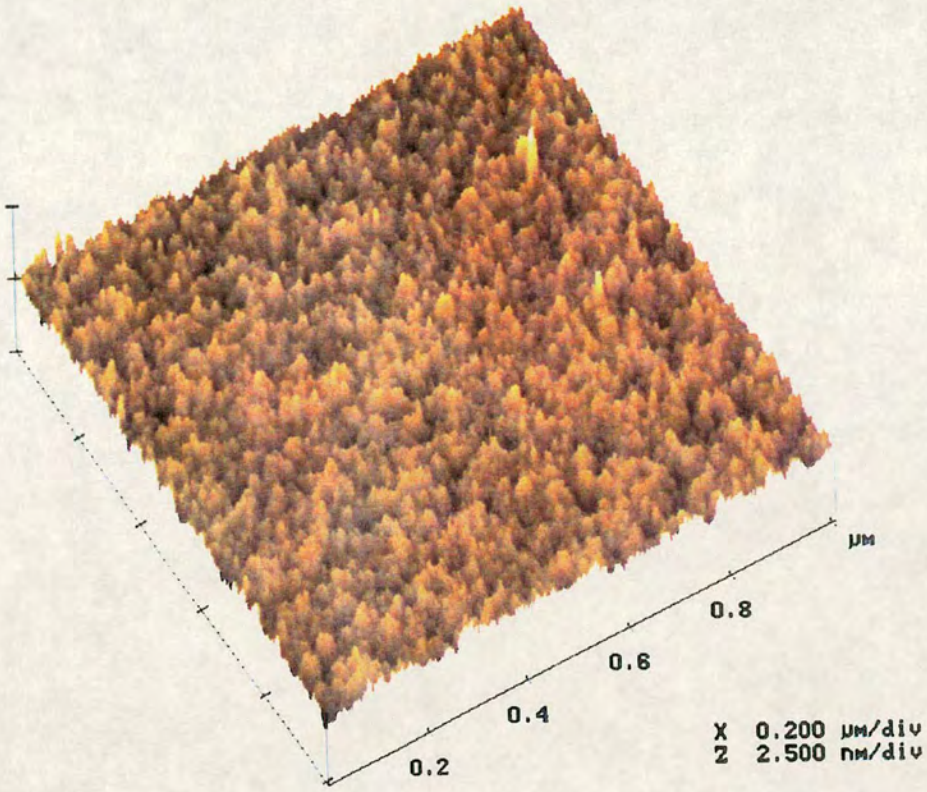


Figure 7.32 Topographic AFM image of the silicon surface with a 1µm by 1µm area and Z range of 5 nm for the 1e14cm^{-2}, 10KeV sample.

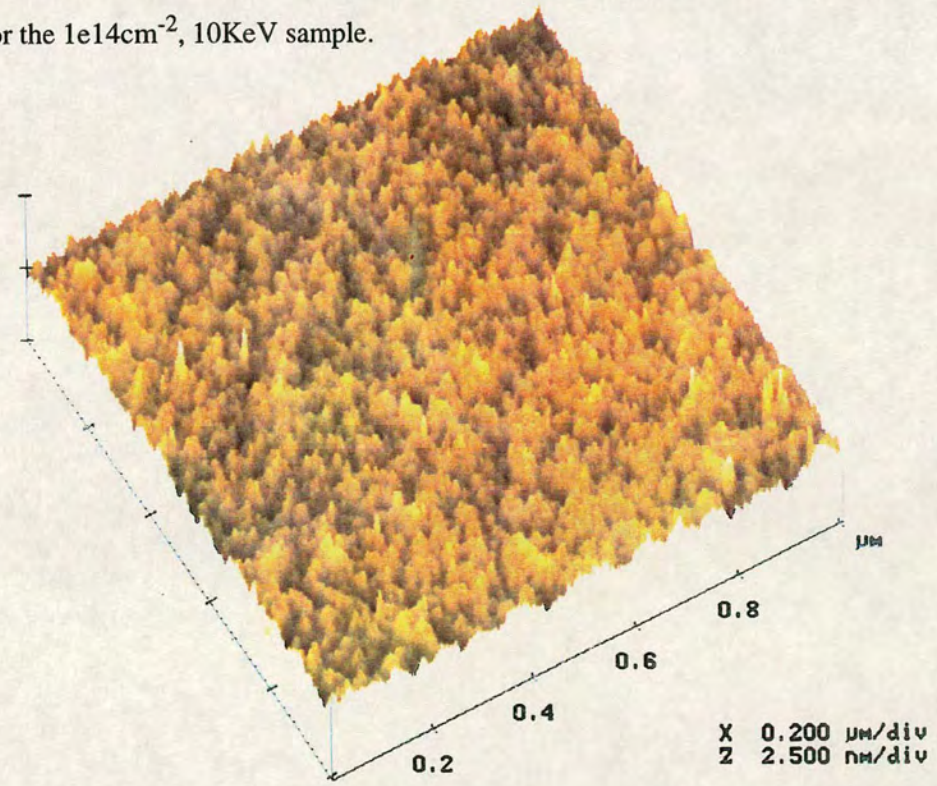


Figure 7.33 Topographic AFM image of the silicon surface with a 1µm by 1µm area and Z range of 5 nm for the 1.2e14cm^{-2}, 10KeV sample.

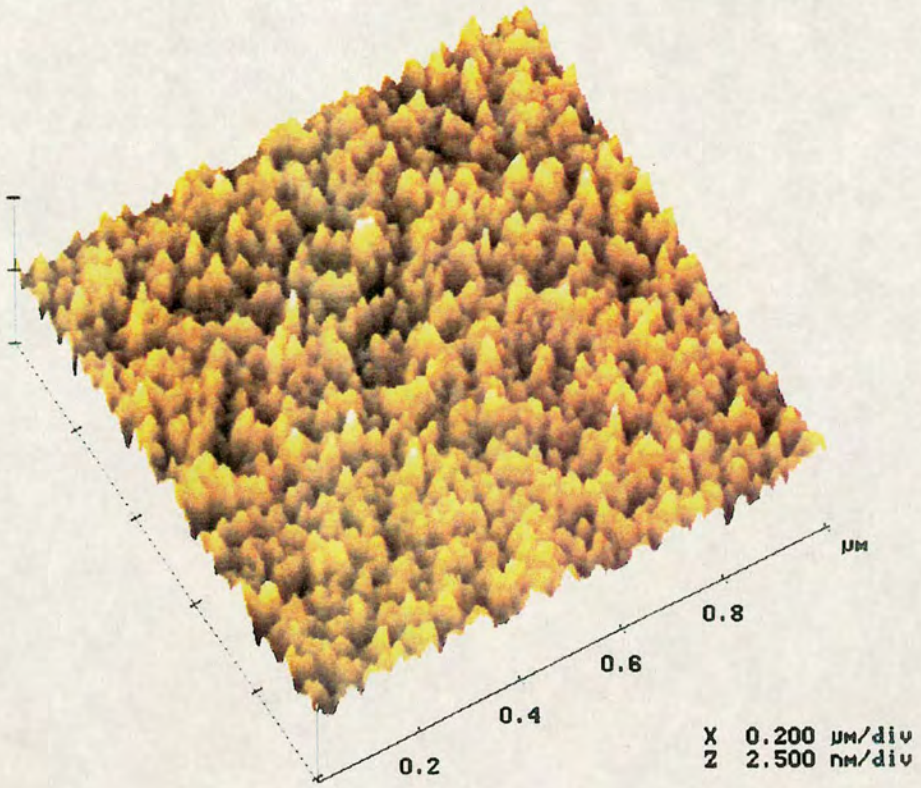


Figure 7.34 Topographic AFM image of the silicon surface with a 1µm by 1µm area and Z range of 5 nm for the 2e14cm^{-2}, 17KeV sample.

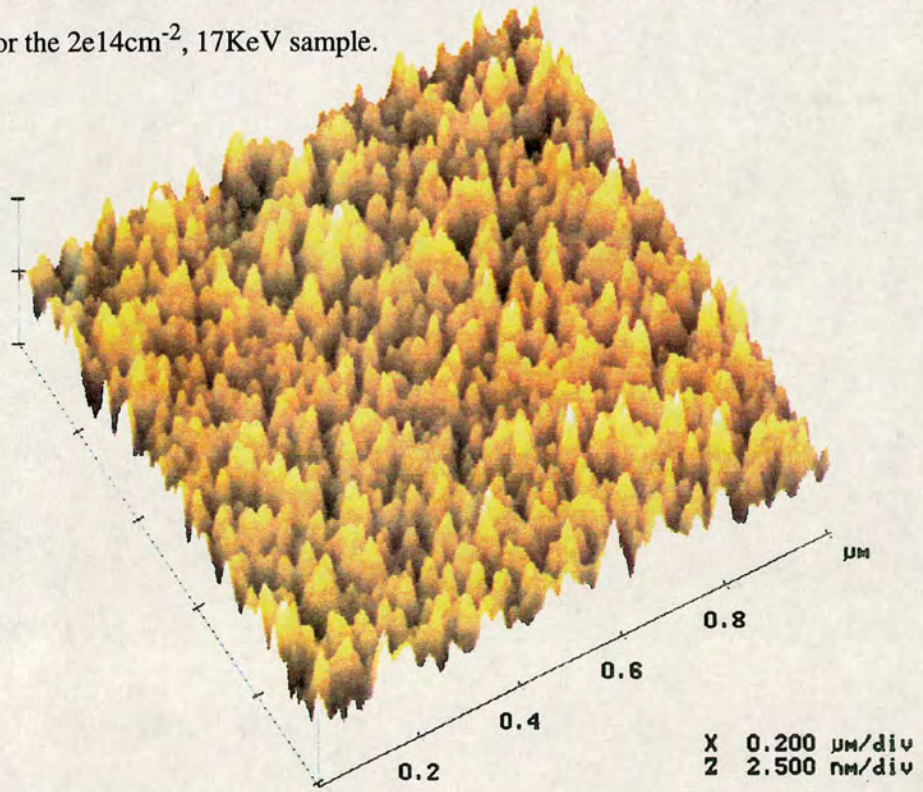


Figure 7.35 Topographic AFM image of the silicon surface with a 1µm by 1µm area and Z range of 5 nm for the 3e14cm^{-2}, 25KeV sample.

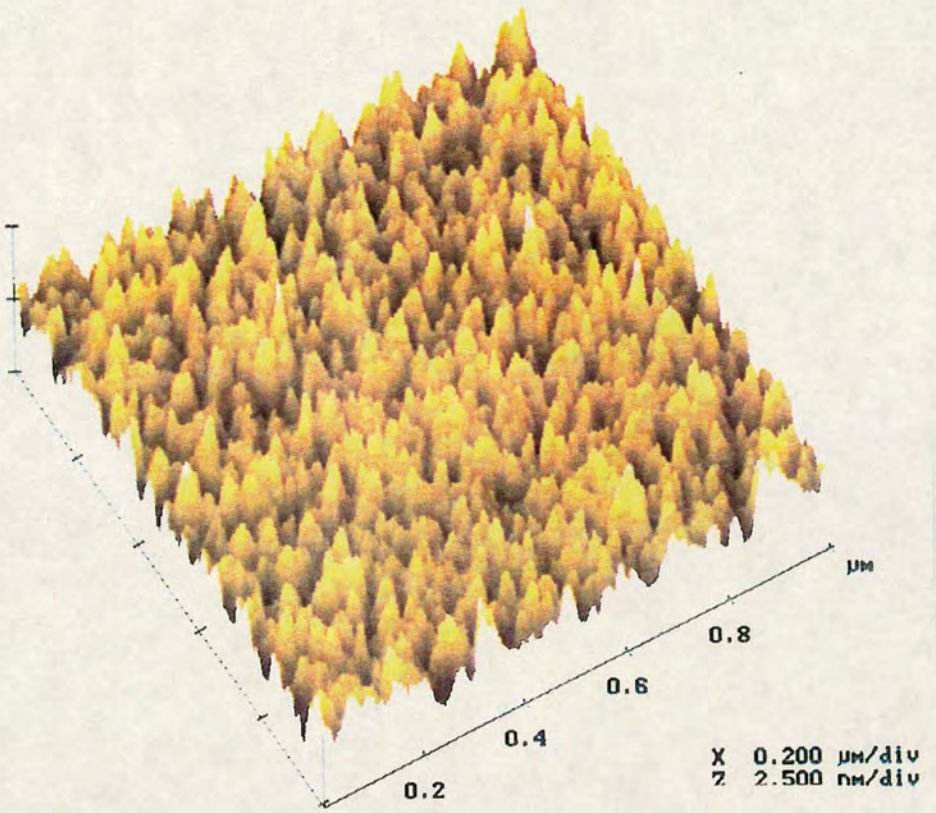


Figure 7.36 Topographic AFM image of the silicon surface with a 1µm by 1µm area and Z range of 5 nm for the 4e14cm^{-2}, 17KeV sample.

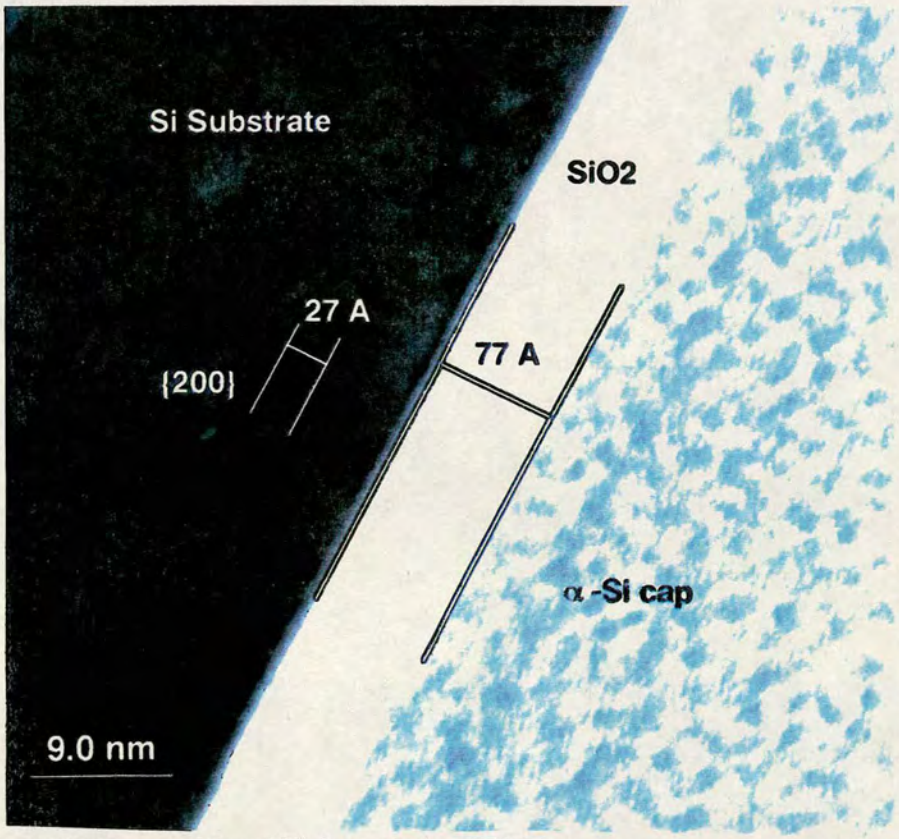


Figure 7.37 Cross-sectional high resolution TEM from the conventional oxide sample.

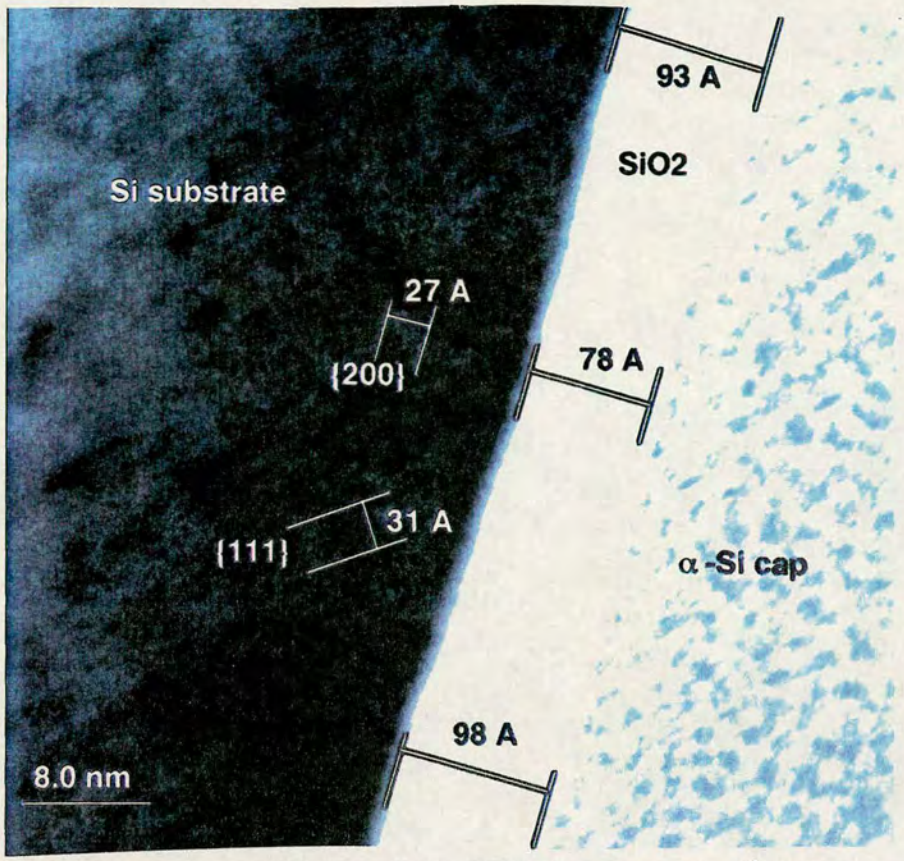


Figure 7.38 Cross-sectional high resolution TEM from the 3e14cm^{-2}, 25KeV sample.

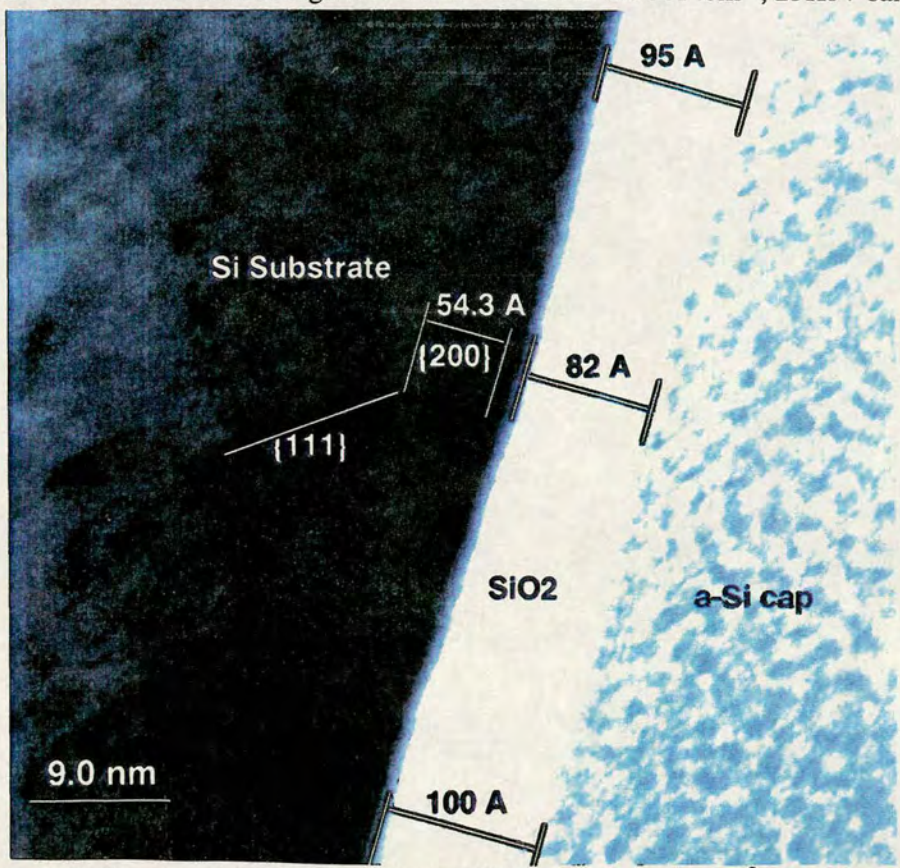


Figure 7.39 Cross-sectional high resolution TEM from the 4e14cm^{-2}, 17KeVsample.

7.5 Conclusions

In this chapter, the performance and reliability characteristics of MOS devices fabricated using a 0.5μm CMOS process have been used to determine the feasibility of the reoxidised molecular nitrogen implanted silicon technique. The splits were modified so that a similar thickness of gate oxide was grown to assess the difference between the splits without the confounding problem of short channel effects. For the range of doses used in this work, the implanted nitrogen does behave like a n-type dopant such that the threshold voltages for both NMOS and PMOS devices change. This result has implications for the integration of the nitrogen implanted silicon technique in terms of threshold voltage matching of the complementary MOS devices. Measurements of the general device characteristics showed that there was a reduction in drive current with increasing molecular nitrogen dose in the silicon for both NMOS and PMOS devices, which was not accounted for by the differences in threshold voltage. Subsequent channel mobility measurements from low to high transverse fields showed that there is increased mobility reduction with nitrogen dose over all transverse fields in both NMOS and PMOS devices. Reverse biased diode breakdown measurements showed no dependence on nitrogen dose which implies that the mobility reduction is not due to ion implant damage.

Dielectric breakdown measurements on large arrays showed that the gate oxide is more susceptible to breakdown with increased nitrogen dose at the polysilicon gate to source/drain periphery. A break point in the gate oxide reliability at this gate edge was shown by the use of antenna structures to be between the 1.2e14cm^{-2},10KeV and 2e14cm^{-2},17KeV molecular nitrogen dose splits. For the splits investigated in this work, the acceptable (100%) breakdown fields up to 8MV/cm for both the NMOS and PMOS polysilicon antenna structures is limited up to the 100:1 antenna ratio. Since the design rules for CMOS5 require acceptable TZDB for up to 30:1 antenna ratio, these results are good for single transistors. The large area capacitors however, show unacceptable breakdown for all splits. While a known gate oxide defectivity mechanism was unintentionally incorporated into this experiment, there is a cause for concern since the large area breakdown is a function of nitrogen content in the silicon.

The hot carrier stress results and boron penetration study show that gate oxides behave some what similarly to oxides grown by other nitridation methods. Boron penetration is reduced and the interface trap limited hot carrier lifetime is increased with increased nitrogen content in the silicon. For the 4e14cm^{-2},17KeV case the hot carrier limited lifetime is almost doubled compared to conventional oxide. This result is in contrast to a 5 times improvement of the CMOS5 ROXNOX gate oxide compared to conventional oxide [172].

An investigation into the cause of the reduced channel mobility and oxide reliability using

atomic force microscopy showed that there is an increased oxide and silicon substrate roughening with nitrogen content in the silicon. TEM's of some of the wafers indicate there is a localised variation in oxidation rate which causes the oxide to be significantly thinner in some areas. These results are explained by postulating that there is a non-uniform distribution of the nitrogen at the silicon surface during oxidation which results in localised oxidation rate variations. The effect of increasing the nitrogen content in the silicon is to increase the localised variation in oxidation rate.

For the application of molecular nitrogen implanted silicon for the growth of different thicknesses of oxide there is seen to be a trade off between the oxide thickness differential and the performance and reliability of the devices. There is a breakpoint between the $1.2e14cm^{-2}$ and the $2e14cm^{-2}$ nitrogen dose splits which determines the dose of nitrogen in the silicon irrespective of energy, in terms of the gate oxide reliability. This range of nitrogen dose is important since it would produce an oxide thinning of 2/3 required to grow both 150Å and a 100Å oxides for 3.3V and 5V applications respectively.

Chapter 8

Conclusions and Further Work

The purpose of the work in this thesis was to determine the feasibility of using molecular nitrogen ion implantation into silicon to selectively retard the oxidation rate so that it is possible to simultaneously grow different thicknesses of gate oxide on a microelectronic circuit. The approach taken in this feasibility study was to determine the performance and reliability characteristics of gate oxides grown on silicon with various low doses of molecular nitrogen by the integration of this technique into an existing 0.5μm CMOS process. In order to easily implement this technique into the CMOS5 process, this work did not include a thermal step after implantation of the nitrogen. Furthermore, to eliminate the confounding effects of short channel effects, the gate oxides were all grown to ~100Å for the different implant conditions. While this did enable a comparison of nitrogen implanted splits with conventional oxide, it did not evaluate the simultaneous growth of different thicknesses of gate oxide. One reason for this was that a mask was not made to allow the selective implantation of molecular nitrogen. From the results of this work however, the next step to evaluate this technique would be to selectively implant the nitrogen and then grow the different thicknesses of gate oxide.

Taking the precondition of growing the same thickness of gate oxide with differing amounts of nitrogen in the silicon, the oxide growth characteristics under different low dose molecular nitrogen implant conditions were studied for the first time. The oxide growth rate was seen to be sharply inhibited with increasing molecular nitrogen dose between $10^{14}cm^{-2}$ and $10^{15}cm^{-2}$ for the range of energies studied. The absence of nitrogen diffusion through an oxide and the evidence of nitrogen pile up at the silicon-silicon oxide interface indicate that the oxidation is reaction rate limited over the relevant range of gate oxide thicknesses. This result is important since it explains why the oxidation characteristics are determined only by the dose of nitrogen in the silicon (for relatively shallow implants) which piles up to the silicon-silicon oxide interface almost instantaneously during oxidation. For heavy molecular nitrogen doses ($>10^{15}cm^{-2}$) the oxidation is inhibited for a long period and then the nitrogen is evaporated from the silicon surface to a concentration that allows oxidation to proceed. The thermal cycles for these heavy doses is such that they are not practical for submicron CMOS applications and so there is a constraint on the maximum concentration of nitrogen at the silicon-silicon oxide interface. For the application to mixed thicknesses of gate oxide on

a circuit, there is a potential benefit of increased nitrogen at the silicon-silicon oxide interface with a thinner gate oxide, this could mean that as the gate oxide is thinned the hot carrier robustness could be improved.

The results of MOS device measurements for performance and reliability aspects relevant to the application of this technique to ULSI circuits showed that there is a trade-off with the degree of oxide thinning. As the nitrogen content in the silicon is increased, the mircoroughness of the gate oxide that is subsequently grown is increased. This results in a decrease in channel mobility over all transverse electric fields and a deterioration in the oxide breakdown distributions. While the oxide breakdown for single transistors was acceptable, a general defectivity problem clouded the results of oxide breakdown for large capacitor arrays. There were indications however that the oxide breakdown showed a consistent deterioration between the $1.2e14cm^{-2}$ and the $2e14cm^{-2}$ molecular nitrogen dose splits compared to conventional oxide. This range of doses is important as it would produce the oxide thinning ratio of 2/3 at $2e14cm^{-2}$ required to grow a 150Å and 100Å gate oxide for 5V and 3.3V power supply CMOS technologies. Extrapolation of the oxide breakdown results at the N_2^+ dose of $2e14cm^{-2}$ for a single transistor to the large area SRAM cell capacitor using a negative binomial model (Equation 4.16) with a clustering coefficient of 0.3 [177], gives a rate of capacitor failure of ~3% at 5MV/cm for yield screening and ~17% between 5 and 8MV/cm for reliability (burn-in) screening. These rates of failure could be tolerated in manufacturing but the process margin is small and hence this technique would have to be improved mainly for gate oxide reliability reasons. The degredation of mobility due to nitrogen incorporation results in a 8% reduction in saturation current. While this would reduce the operating speed of a circuit, it would only be significant for high performance microprocessor circuits. Since the localised oxide thinning implies that the nitrogen is inhomogeneously distributed during oxidation, one possible solution could be to anneal the silicon after N_2^+ implantation and hopefully distribute the nitrogen more uniformly. The MOS threshold and punchthough channel implants would have to be re-engineered to accommodate the post N_2^+ implant anneal.

The improvements of reduced boron penetration and improved hot carrier limited lifetimes with increased nitrogen content are not significant for the $2e14cm^{-2}$ split. Since other methods of gate oxide nitridation show greater improvements in hot carrier lifetime and boron penetration, this technique of nitrogen implantation in the silicon prior to oxidation is

not recommended due to the high doses required and the gate oxide reliability problem. Instead, this technique of nitrogen implantation into the silicon is proposed for only the application to the simultaneous growth of different gate oxide thicknesses. The potential benefits of this technique are commercially attractive. The ability to interface low voltage high speed transistors in the core of a microprocessor with a thicker gate oxide for the input driver transistors which see the full high voltage from the peripheral chips, would allow the faster progression to higher performance systems without the need to wait on the advancement of the power supply voltage reduction of chip to chip interconnects.

The main objective of this work has been met, namely this technique has been applied to an existing CMOS process to show (albeit in an indirect way) that it is possible to use this technique to simultaneously grow gate oxide of different thickness on a single ULSI microchip. In addition, the limitations of this technique have been evaluated in terms of MOS device performance and reliability and suggestions for improvements to this process are made.

REFERENCES

[1] B. Hoefflinger and G. Zimmer, "New CMOS Technologies", 10th European Solid-state Device Research Conf., Sept., 1980.

[2] C. Hu, "Future CMOS Scaling and Reliability", Proc. IEEE, May, p682, 1993.

[3] H. Iwai et al., "Analysis of Velocity Saturation and Other Effects in Short-Channel MOS Transistor Capacitances", IEEE Transactions on Computer Aided Design, vol. CAD-6, pp.173-184, 1987.

[4] G. Baccarani et al., "Generalized Scaling Theory and its application to a 1/4 Micrometer MOSFET Design", IEEE Transactions on Electron Devices, ED-31, p452, 1984.

[5] J. R. Pfiester et al., "The Effects of Boron Penetration on P+ Polysilicon Gated PMOS Devices", IEEE Trans. Elect. Dev., ED-37, pp. 1842-1851, 1990.

[6] J. E. Chung et al., "Performance and Reliability Design Issues for Deep-Submicrometer MOSFET's", IEEE Trans. Elect. Dev., ED-38, p545, 1991.

[7] P. Olivo et al., "High Field Induced Degredation in Ultra-Thin SiO_2 Films", IEEE Trans. Elect. Dev., ED-35, pp.2259-2267, 1988.

[8] K. M. Cuy, "Design Considerations Bring Unity to a Mixed-Voltage World", EDN, Feb. 2nd, p115, 1995.

[9] K. K. O et al., "Integration of Two Different Gate Oxide Thicknesses in a 0.6μm Dual Voltage Mixed Signal CMOS Process", IEEE Trans. Elec. Dev., ED-42, pp. 190-192, 1995.

[10] B.S. Doyle, H.R. Soleimani and A. Philipossian, "Simultaneous Growth of Different Thickness Gate Oxides in Silicon CMOS Processing", IEEE Electron Device Letters, EDL-16, p 301, 1995.

[11] S. Rigo et al., "Silica Films on Silicon", in Instabilities in Silicon Devices - Vol. 1 Chap. 1, Eds. G. Barbottin and A. Vapaille, Elsevier, Amsterdam, 1986.

[12] B.E. Deal and A.S. Grove, General Relationship for the Thermal Oxidation of Silicon", J. Applied Physics, vol.36, p 3770, 1965.

[13] H. Z. Massoud, J.D. Plummer and E.A. Irene, J. Electrochemical Soc., vol. 132, p2693, 1985.

[14] A. Reisman et al., J. Electron Materials, vol 16, p45, 1987.

[15] Y. Mii et al., "An Ultra-Low Power 0.1μm CMOS", Proc. Symp. VLSI Tech., p.9, 1994.

[16] A.C. Adams et al., J. Electrochemical Soc., vol. 127, p1787, 1980.

[17] Y.Z. van der Meulin, J. Electrochem. Soc., vol. 119, p530, 1972.

[18] C.M. Osburn et al., "Silicon Gate Oxide Thickness Uniformity during HCl Oxidation", J. Electrochem. Soc., vol. 138, p268, 1991.

[19] A. Reisman et al., J. Electrochem. Soc., vol. 132, p284, 1990.

[20] R. Pong, "LPCVD - Are Vertical Reactors the Answer?", Microelectronics Mtg. and Test., July, p1, 1990.

[21] G. Gould and A.E. Irene, "The Influence of Silicon Surface Cleaning Procedures on Silicon Oxidation", J. Electrochem. Soc., vol. 134, p1031, 1987.

[22] M. Meuris et al., "Investigating Techniques to Improve Gate-Oxide Integrity", Microcontamination, May, p31, 1991.

[23] D. Dimetrius et al., "Effects of HF-Last Clean Process Sequence on Gate-Oxide Quality", Abs. Mtg. Electrochem. Soc., p 473, Fall 1993.

[24] C. Werkhoven et al., "Wet and Dry HF-Last Cleaning Process for High-Integrity Gate Oxides", Tech. Dig. IEDM, p633, 1992.

[25] B.E. Deal., "Standardized Terminology for Oxide Charges Associated with Thermally Oxidized Silicon", IEEE Trans. Elect. Dev., ED-27, p606, 1980.

[26] B.E. Deal, J. Electrochem. Soc., vol. 121, 198C, June 1974.

[27] B.E. Deal, J. Electrochem. Soc., vol. 114, p266, 1967.

[28] E.H. Nicollian and J.R. Brews, MOS Physics and Technology, Wiley-Interscience, New York, 1982.

[29] C.E. Young, "Extended Curves of the Space Charge, Electric Field, and Electrostatic Potential Inside a Semiconductor", J. Appl. Phys., vol. 32, p.329, 1961.

[30] S.M. Sze, Physics of Semiconductor Devices, 2nd. Ed., Wiley & Sons, New York, 1981.

[31] Y.P. Tsividis, Operation and Modeling of the MOS Transistor", McGraw-Hill, New York, 1987.

[32] R.F. Pierret and J.A. Shields, "Simplified Long-Channel MOSFET Theory", Solid-State Electronics, vol. 26, p.143, 1983.

[33] R.S. Muller and T.I. Kamis, Device electronics for Integrated Circuits", 2nd. Ed., New York, John Wiley & Sons, 1986.

[34] S.M. Sze, Semiconductor Devices: Physics and Technology, John Wiley & Sons, New York, 1985.

[35] D.K. Schroder, Advanced MOS Devices. modular Series on Solid State Devices, Reading, Mass., Addison-Wesley, 1987.

[36] E.S. Yang, Microelectronic Devices, McGraw-Hill, New York, 1988.

[37] L.D. Yau, "A Simple Theory to Predict the Threshold Voltage of Short-Channel IGFETs", Solid-State Electronics, pp. 1059-1063, 1974.

[38] L.A. Akers and J.J. Sanchez, "Threshold Voltage Models of Short, Narrow, and Small Geometry MOSFETs: A Review", Solid-State Electronics, pp. 621-641, 1982.

[39] Y. Nissan-Cohen et al., "Measurements of Fowler-Nordheim Tunnelling Currents in MOS Structures under Charge Trapping Conditions", Solid-State Electronics, p717, 1985.

[40] J.R. Brews, "The Submicron MOSFET", Chap. 3, High Speed Semiconductor Devices, Ed. S.M. Sze, Wiley Interscience, New York, 1990.

[41] K. Yambe and K. Taniguchi, "Time-Dependent Dielectric Breakdown of Thin Thermally grown SiO_2 Films", IEEE Trans. Elect. Dev., ED-32, p.423, 1985.

[42] E. Harari, "Dielectric Breakdown in Electrically Stressed Thin Films of SiO_2 Films", J. Appl. Phys., vol. 49, no.4, p.2478, 1978.

[43] B. Ricco et al., "Novel Mechanism for Tunnelling and Breakdown of Thin SiO_2 Films", Phys. Rev. Lett., vol. 51, p.1795, 1983.

[44] Y. Nissan-Cohen and T. Gorczyca, "The Effect of Hydrogen on Trap Generation, Positive Charge Trapping, and Time-Dependent Dielectric Breakdown of Gate Oxides", IEEE Elect. Dev. Lett., p.287, 1988.

[45] I.C. Chen, S. Holland and C. Hu, "Electrical Breakdown of Thin Gate and Tunnelling Oxides", IEEE Trans. Elect. Dev., p.413, 1985.

[46] C. Hu, "Thin Oxide Reliability", Tech. Dig. IEDM, p.368, 1985.

[47] I.C. Chen, S. Holland and C. Hu, "Oxide Breakdown Dependence on Thickness and Hole Current-Enhanced Reliability of Ultra-Thin Oxides", Tech. Dig. IEDM, p.660, 1986.

[48] K.F. Schuegraf and C. Hu, "Hole Injection SiO_2 Breakdown Model for Very Low Voltage Lifetime Extrapolation", IEEE Trans. Elect. Dev., ED-41, p.761, 1994.

[49] J.C. Lee, I.-C. Chen and C. Hu, "Modeling and Characterization of Gate Oxide Reliability", IEEE Trans. Elect. Dev., ED-35, p.2268, 1988.

[50] S. Verhaverbeke et al., "The Effect of Metallic Impurities on the Dielectric Breakdown of Oxides and Some Ways of Avoiding Them", Tech. Dig. IEDM, p.71, 1991.

[51] M. Liehr, G.B. Bronner and J.E. Lewis, "Stacking-Fault-Induced Defect Creation in SiO_2 on Si (100)", appl. Phys. lett., vol. 52, p.1892, 1988.

[52] M. Miyashita et al., "Dependence of Surface Microroughness of CZ, FZ and Epi Wafers on Wet Chemical Processing", J. Electrochem. Soc., p.2133, 1992.

[53] K. Shiozaki et al., "Improvement of Stress-Induced Surface Microroughness for Highly Reliable Gate Oxide", Ext. Abs. Electrochem. Soc. Mtg., Abs. No. 135, p.207, Spring 1994.

[54] H. Uchida et al., "The Effect of Oxide Charges at LOCOS Isolation Edges on Oxide Breakdown", IEEE Trans. Elect. Dev., ED-40, p.1818, 1993.

[55] E. Kooi, The Invention of LOCOS, IEEE Press, New York, 1991.

[56] C.T. Gabriel, "Gate Oxide Damage from Polysilicon Etching", J. Vac. Sci. Tech. B vol. 9(2), p.370, 1991.

[57] C.T. Gabriel and J.P. McVittie, "How Plasma Etching Damages Thin Gate Oxides", Solid-State Technology, p. 81, June 1992.

[58] H. Shin et al., "Plasma Etching Charge-Up Damage to Thin Oxides", Solid-State Technology, p.29, August 1993.

[59] D. Jackson, DEC Internal Memo., 12th July, 1989.

[60] C.M. Osburn and D.W. Ormond, J. Electrochem. Soc., vol. 121, p.526, 1972.

[61] A. Berman, "Time-Zero Dielectric Reliability Test by a Ramp Method", Proc. IRPS, p.204, 1981.

[62] K.F. Schuegraf, C.C. King and C. Hu, "Impact of Polysilicon Depletion in Thin Oxide MOS Technology", Proc. Symp. VLSI Tech., p.86, 1993.

[63] D.L. Crook, "Method of Determining Reliability Screens for Time-Dependent Dielectric Breakdown", Proc. IRPS, p.1, 1979.

[64] R. Moazzami and C. Hu, "Projecting Gate Oxide Reliability and Optimizing Reliability Screens", IEEE Trans. Elect. Dev., ED-37, p.1643, 1990.

[65] S. Fang and J.P. McVittie, "Thin-Oxide Damage from Gate Charging During Plasma Processing", IEEE Elect. Dev. Lett., p.288, 1992.

[66] S. Fang, S. Murakawa and J.P. McVittie, "A New Model for Thin Oxide Degredation from Wafer Charging in Plasma Etching", Tech. Dig. IEDM, p.61, 1992.

[67] P.H. Singer, "Evaluating Plasma Etch Damage", Semiconductor International, p.78, May 1992.

[68] W.M. Greene, J.B. Kruger and G. Kooi, "Magnetron Etching of Polysilicon: Electrical Damage", J. Vac. Sci. Tech. B vol. 9(2) p.1638, 1989.

[69] R. de Werdt et al., Tech. Dig. IEDM, p.532, 1987.

[70] H. Shin, Z.-J. Ma and C. Hu, "Impact of Plasma Charging Damage and Diode Protection on Scaled Thin Oxide", Tech. Dig. IEDM, p.467, 1993.

[71] T.H. Ning et al., "1μm MOSFET VLSI Technology: Part IV - Hot Electron Design Constraints", IEEE Trans. Elect. Dev., ED-26, p.346, 1979.

[72] P.K. Ko, R.S. Muller and C. Hu, "A Unified Model for Hot-electron Currents in MOSFETs", Tech. Dig. IEDM, p.600, 1981.

[73] J.M. Pimbley et al., Advanced CMOS Process Technology, vol. 19, VLSI Electronics Microstructure Science, Ch. 5, "Reliability", Academic Press, San Diego, 1989.

[74] C. Sodini, P.K. Ko and J.L. Moll, "The Effect of High Fields on MOS Device and Circuit Performance", IEEE Trans. Elect. Dev., ED-31, p.1386, 1984.

[75] P.K. Chatterjee, Tech. Dig. IEDM, p.14, 1979.

[76] E. Takeda et al., "Submicrometer MOSFET Structures for Minimizing Hot-Carrier Generation", IEEE Trans. Elect. Dev., ED-29, p.611, 1982.

[77] T.Y. Chan, P.K. Ko and C. Hu, "A Simple Method to Characterize Substrate Current in MOSFETs", IEEE Elect. Dev. Lett., p.505, 1984.

[78] M.L. Woods, "MOS VLSI Reliability and Yield Trends", Proc. IEEE, p.1715, 1986.

[79] P.Yang and S. Aur, "Modeling of Device Lifetime due to Hot Carrier Effects", Proc. Symp. VLSI Tech., p.227, 1985.

[80] E. Takeda and N. Suzuki, "An Empirical Model for Device Degredation due to Hot Carrier Injection", IEEE Elect. Dev. lett., p.111, 1983.

[81] C. Hu et al., "Hot-Electron-Induced MOSFET degredation - Model, Monitor and Improvement", IEEE Trans. Elect. Dev., ED-32, p.375, 1985.

[82] K.R. Hoffmann et al., "Hot-electron and Hole-Emission Effects in Short N-MOSFETs", IEEE Trans. Elect. Dev., ED-32, p.691, 1985.

[83] B.S. Doyle et al., "Dynamic Channel Hot-Carrier Degredation of MOS Transistors by Enhanced Electron-Hole Injection into the Oxide", IEEE Elect. Dev. Lett., p.237, 1987.

[84] B.S. Doyle et al., "The Generation and Characterization of Electron and Hole Traps Created by Hole Injection During Low Gate voltage Hot Carrier Stressing of N-MOS Transistors", IEEE Trans. Elect. Dev., ED-37, p.1869, 1990.

[85] B.S. Doyle et al., "Interface State Creation and Charge Trapping in the Medium-to-High Gate Voltage During Hot Carrier Stressing of N-MOS Transistors", IEEE Trans. Elect. Dev., ED-37, p.744, 1990.

[86] M. Yoshida et al., "Increase in Resistance to Hot Carriers in Thin Oxide MOSFETs", Tech. Dig. IEDM, p.200, 1988.

[87] P. Woerlee et al., "The Impact of Scaling on Hot-Carrier Degredation and Supply Voltage of Deep Submicron NMOS Transistors", Tech. Dig. IEDM, p.537, 1991.

[88] H. Hazama, M. Iwase and S. Takagi, "Hot-Carrier Reliability in Deep Submicrometer MOSFETs", Tech. Dig. IEDM, p.569, 1990.

[89] K.R. Mistry and B.S. Doyle, "AC Verses DC Hot Carrier Degredation in N-Channel MOSFETs", IEEE Trans. Elect. Dev., ED-40, p.96, 1993.

[90] Y. Hiruta et al., "Impact of Hot-electron Trapping of Half-Micron PMOSFETs with p+ Poly-Si Gate", Tech. Dig. IEDM, p.718, 1986.

[91] R. Woltjer and G.M. Paulzen, "Oxide Charge Generation During Hot-Carrier Degredation in PMOSTs", Tech. Dig. IEDM, p.713, 1993.

[92] M. koyonagi et al., "Investigation and Reduction of Hot-Electron Induced Punchthrough (HEIP) Effect in Submicron PMOSFETs", Tech. Dig. IEDM, p.722, 1986.

[93] H. Mikoshiba, "Comparison of Drain Structures in N-Channel MOSFETs", IEEE Trans. Elect. Dev., ED-33, p.140, 1986.

[94] M. Koyonagi, H. Kaneko and S. Shinizu, "Optimum Design of n+-n-Double-Diffused Drain MOSFET to Reduce Hot-Carrier Emission", IEEE Trans. Elect. Dev., ED-32, p.562, 1985.

[95] S. Orura et al., "Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate FET", IEEE Trans. Elect. Dev., ED-27, p.1359, 1980.

[96] S.H. Dhong and E.J. Petrillo, "Sidewall Spacer Technology for MOS and Bipolar Devices", J. Electrochem. Soc., p.389, 1986.

[97] K. Marayam, J.C. Lee and C. Hu, "A Model for the Electric Field in Lightly-Doped Drain Structures", IEEE Trans. Elec. Dev., ED-34, p.1509, 1987.

[98] R. Izawa and E. Takeda, "The Impact of n- Drain Length and Gate/Drain-Source Overlap on Submicrometer LDD Devices for VLSI", IEEE Elect. Dev. Lett., p.480, 1987.

[99] T. Mizuno et al., "A New Degredation Mechanism of Current Drivability and Reliability of Asymmetrical LDD MOSFETs", Tech. Dig. IEDM, p.250, 1985.

[100] G. Krieger et al., "Shadowing Effects Due to Tilted Arsenic source/Drain Implant", IEEE Trans. Elect. Dev., ED-36, p.1248.

[101] B.S. Doyle, C. Bergonzoni and A. Boudou, "The influence of Gate Edge Shape on the Degredation in Hot-Carrier Stressing of N-Channel Transistors", IEEE Elect. Dev. Lett., p.363, 1991.

[102] T.Y. Chan et al., "Effects of the Gate-to-Drain/Source Overlap on MOSFET Characteristics", IEEE Elect. Dev. Lett., p.326, 1987.

[103] F. Hsu and K.-Y. Chiu, "Evaluation of LDD MOSFETs Based on Hot-Electron-Induced Degredation", IEEE Elect. Dev. Lett., p.162, 1984.

[104] G. Krieger et al., "Moderately Doped NMOS (M-LDD) - Hot Electron and Current Drive Optimization", IEEE Trans. Elect. Dev., ED-38, p.121, 1991.

[105] J.J. Sanchez, K.K. Hsueh and T.A. DeMassa, "Drain-Engineered Hot-Electron-Resistant Device Structures: A Review", IEEE Trans. Elect. Dev., ED-36, p.1125, 1989.

[106] C. Wei et al., "Buried and Graded/Buried LDD Structures for Improved Hot-Electron Reliability", IEEE Elect. Dev. Lett., p.380, 1986.

[107] C. Codella and S. Ogura, "Halo doping Effects in Submicon DI-LDD Design", Tech. Dig. IEDM, p.230, 1985.

[108] T. Hori et al., "Deep Submicrometer LATID Technology", IEEE Trans. Elect. Dev., ED-39, p2312, 1992.

[109] T. Hori et al., "High Performance Dual Gate CMOS Utilizing a Novel Self-Aligned Pocket Implantation (SPI) Technology", IEEE Trans. Elect. Dev., ED-40, p1673, 1993.

[110] T. Huang et al, "A New LDD Transistor with Inverse-T Gate Structure", IEEE Elect. Dev. Lett., p.151, 1987.

[111] R. Izawa, T. Kure and E. Takeda, "Impact of the Gate-Overlapped Device (GOLD) for Deep Submicrometer VLSI", Tech. Dig. IEDM, p.38, 1987.

[112] T. Mizuno et al. "Si3N4/SiO2 Spacer Induced High Reliability in LDD MOSFET and its Simple Degredation Model", Tech. Dog. IEDM, p.234, 1988.

[113] T. Mizuno et al., "High Dielectric LDD Spacer Technology for High Performance MOSFET using Gate-Fringing Field Effects", Tech. Dig. IEDM, p.613, 1989.

[114] T. Hori et al., "A New Submicron MOSFET with LATID (LArge-Tilt-angle, Implanted Drain) Structure", Proc. Symp. VLSI Tech., p.15, 1988.

[115] T. Hori et al., "High Carrier Velocity and Reliability of Quarter-Micron SPI (Self-Aligned Pocket Implantation) MOSFETs", Tech. Dig. IEDM, p.699, 1992.

[116] T. Buti et al., "Asymmetrical Halo Source-GOLD Drain (HS-GOLD) Deep Sub-half Micron N-MOSFET Design for Reliability and Performance", Tech. Dig. IEDM, p.617, 1989.

[117] R. B. Fair and R.C. Sun, "Threshold Voltage Stability in MOSFETs due to Channel Hot-Hole Emission", IEEE Trans. Elect. Dev., ED-28, p.83, 1981.

[118] J. Mitsuhashi, S. Nakao and T. Matsukawa, "Mechanical Stress and Hydrogen Effects on Hot Carrier Injection", Tech. Dig. IEDM, p.387, 1986.

[119] M.T. Takagi, I. Yoshii and K. Hashimoto, "Characterization of Hot-Carrier-Induced Degredation of MOSFETs Enhanced by H2O diffusion for Multilevel Interconnection Processing", Tech. Dig. IEDM, p.703, 1992.

[120] N. Lifshitz and G. Smolinsky, "Hot-Carrier Aging of the MOS Transistor in the Presence of Spin-on Glass at the Interlevel Dielectric", IEEE Elect. Dev. Lett., p.140, 1991.

[121] Y. Ohji et al., "Effects of Minute Impurities (H, OH, F) on SiO2/Si Interface as Investigated by Nuclear Resonant Reaction and Electron Spin Resonance", IEEE Trans. elect. Dev., ED-37, p.1635, 1990.

[122] F. Hsu and K. Chiu, "Effects of Device Processing on Hot-Electron Induced Device Degredation", Proc. Symp. VLSI Tech., p.108, 1985.

[123] F.C. Hsu, J. Hui and K.Y.Chiu, "Effect of Final annealing on Hot-Electron-Induced MOSFET Degredation", IEEE Elect. Dev. Lett., p.369, 1985.

[124] X.-Y. Li et al., "Plasma-Damaged Oxide Reliability Study Correlating Both Hot-Carrier Injection and Time-Dependent Dielectric Breakdown", IEEE Elect. Dev. Lett., p.91, 1993.

[125] K.R. Mistry, B.J. Fishbein and B.S. Doyle, "Effect of Plasma-Induced Charging Damage on N-Channel and P-Channel MOSFET Hot Carrier Reliability", Proc. IRPS, p.42, 1994.

[126] X.-Y. Li et al., "Degraded CMOS Hot-Carrier Lifetime - Role of Plasma Etching Induced Charging Damage and Edge Damage", Proc. IRPS, p.260, 1995.

[127] X.-Y. Li et al., "Effect of Poly Etch on Effective Channel Length and Hot Carrier Reliability in Submicon Transistors", IEEE Elect. Dev. Lett., p.285, 1994.

[128] C. Wong et al., "Doping of n+ and p+ Polysilicon Gates in Dual-Gate CMOS Process", Tech. Dig. IEDM, p.238, 1988.

[129] Y. Nishioka et al., "Hot-electron Hardened Si-Gate MOSFET Utilizing F Implantation", IEEE Elect. Dev. Lett., p.141, 1989.

[130] V. Jain et al., "Improved Sub-Micron CMOS Device Performance in CVD Tungsten Silicide", Proc. Symp. VLSI Tech., p.91, 1991.

[131] Y. Nishioka et al., "Dramatic Improvement of Hot-Electron-Induced Interface Degredation in MOS Structures Containing F or Cl in SiO2", IEEE Elect. Dev. Lett., p.38, 1988.

[132] J.A. Nemetz and R.F. Tressler, "Thermal Nitridation of Silicon and Silicon dioxide for Thin Gate Insulators", Solid State Technology, p.209, September 1983.

[133] T. Ito, T. Nozaki and H. Ishikawa, "Direct Thermal Nitridation of Silicon Oxide Films in Anhydrous Ammonia Gas", J. Electrochem. Soc., p.2053, 1980.

[134] T. Kaga and T. Hagiwara, "Short- and Long-Term Reliability of Nitridation Anneals of SiO2 for Ultrathin Dielectrics", IEEE Trans. Elect. Dev., ED-37, p.1836, 1990.

[135] C.-T. Chen et al., "Study of electrical Characteristics on Thermally Nitrided SiO2 Films", J. Electrochem. Soc., p.875, 1984.

[136] S.-T. Chang, N.M. Johnson and S.A. Lyon, "Capture and Tunnel Emission of electrons by Deep Levels in Ultrathin Nitrided Oxides on Silicon", J. Appl. Phys. Lett., p316, 1984.

[137] S.K. Lai, J. Lee and V.K. Dham, "electrical Properties of Nitrided Oxide Systems for use in Gate Dielectrics and EEPROM", Tech. Dig. IEDM, p.180, 1993.

[138] S.S. Wong et al. "Composition and Electrical Properties of Nitrided-Oxide and Re-Oxidized Nitrided Oxide", Proc. Symp. Silicon Nitride Thin Insulating Films, Electrochem. Soc., p.346, 1983.

[139] F.C. Hsu and K.-Y. Chiu, "A Comparative Study of Tunnelling, Substrate Hot-electron and Channel Hot-electron Injection Induced Degredation in Thin-Gate MOSFETs", Tech. Dig. IEDM, p.96, 1984.

[140] T. Hori, H.Iwasaki and K. Tsuji, "Charge-Trapping Properties of Ultrathin Nitrided Oxides Prepared by Rapid Thermal Annealing", IEEE Trans. Elect. Dev., ED-36, p.904, 1988.

[141] T. Hori, H.Iwasaki and K. Tsuji, "Electrical and Physical Properties of Ultrathin Reoxidized Nitrided Oxides by Rapid Thermal Processing", IEEE Trans. Elect. Dev., ED-36, p.340, 1989.

[142] W. Yang, R. Jayaraman and C.G. Sodini, "Optimization of Low-Pressure Nitridation/ Reoxidation of SiO2 for Scaled MOS devices", IEEE Trans. Elect. Dev., ED-35, p.935, 1988.

[143] B.J. Gross, K.S. Krisch and C.G. Sodini, "An Optimized 850 °C Low-Pressure-Furnace Reoxidized Nitrided Oxide (ROXNOX) Process", IEEE Trans. Elect. Dev., ED-38, p.2036, 1991.

[144] H.S. Momose et al., "Very Lightly Nitrided Oxide MOSFETs for Deep-Sub-Micron CMOS Devices", Tech. Dig. IEDM, p.359, 1991.

[145] B.S. Doyle and A. Philipossian, "p-channel Hot-Carrier Optimization of RNO Gate Dielectrics Through the Reoxidation Step", IEEE Elect. Dev. Lett., p.161, 1993.

[146] A. Uchiyama et al., "High Performance Dual-Gate Sub-Halfmicron CMOSFETs with 6 nm-thick Nitrided SiO2 Films in an N2O Ambient", Tech. Dig. IEDM, p.425, 1990.

[147] H. Hwang et al., "Electrical and Reliability Characteristics of Ultrathin Oxynitride Gate Dielectric Prepared by Rapid Thermal Processing in N2O", Tech. Dig. IEDM, p.421, 1990.

[148] H. Fukuda et al. "Novel N2O - Oxynitridation Technology for Forming Highly Reliable EEPROM Tunnel Oxide Films", Tech. Dig. IEDM, p.587, 1991.

[149] H. Hwang et al., "Improved Reliability Characteristics of Submicrometer MOSFETs with Oxynitride Gate Dielectric Prepared by Rapid Thermal Oxidation", IEEE Elect. Dev. Lett., p.495, 1991.

[150] N.S. Saks et al., "Characteristics of Oxynitrides Grown in N2O", Ext. Abs. Mtg. electrochem. Soc., p.251, Spring 1994.

[151] H.G. Pomp et al. "Lightly N2O Nitrided Dielectrics Grown in a Conventional Furnace for EEPROM and 0.25μm CMOS", Tech. Dig. IEDM, p.463, 1993.

[152] H.R. Soleimani, A. Philipossian and B.S. Doyle, "A Study of the Growth Kinetics of SiO2 in N2O", Tech. Dig. IEDM, p.629, 1992.

[153] J. Ahn et al., "High Quality Ultrathin Gate Dielectrics Formation by Thermal Oxidation of Si in N2O", J. Electrochem. Soc., vol. 138, pL.39, 1991.

[154] Z. Liu et al., "Effects of N2O Anneal and Reoxidation on Thermal Oxide Characteristics", IEEE Elect. Dev. Lett., p.402, 1992.

[155] P.J. Tobin et al., "Silicon Oxynitride Formation in Nitrous Oxide (N2O): The Role of Nitric Oxide", Proc. Symp. VLSI Tech., p.51, 1993.

[156] Y. Okada et al., "Gate Oxynitride Grown in Nitric Oxide (NO)", Proc. Symp. VLSI Tech. p.105, 1994.

[157] B. Maiti, "Reoxidized Nitric Oxide(ReoxNO) Process and its Effect on the Dielectric Reliability of the LOCOS Edge", Proc. Symp. VLSI Tech., p.63, 1995.

[158] S. Haddad and M.-S. Liang, "Improvement of Thin-Gate Oxide Integrity Using Through-Silicon-Gate Nitrogen Ion Implantation", IEEE Elect. Dev. Lett., p.58, 1987.

[159] T. Kuroi et al., "Novel NICE (NItrogen Implanted into CMOS Gate Electrode and Source-Drain) Structure for High Reliability and High Performance 0.25μm Dual Gate CMOS", Tech. Dig. IEDM, p.325, 1993.

[160] T. Kuroi et al., "The Effects of Nitrogen Implantation into p+ Polysilicon Gate on Gate Oxide Properties", Proc. Symp. VLSI Tech., p.107, 1994.

[161] S. Shimizu et al., "0.15μm CMOS Process for High Performance and Reliability", Tech. Dig. IEDM, p.67,1994.

[162] S.C. Sun, "Rapid Thermal Chemical Vapor Deposition of In-Situ Nitrogen Doped Polysilicon for Dual Gate CMOS", Proc. Symp. VLSI Tech., p.121, 1995.

[163] W. J. M. J. Josquin and Y. Tamminga, "The Oxidation Inhibition in Nitrogen-Implanted Silicon", J. Electrochem. Soc., vol. 129., p.1803, 1982.

[164] W. J. M. J. Josquin, Nuclear Instruments and Methods, vol. 209/210, p.581, 1983.

[165] T. Enomoto et al., Jpn. J. Appl. Phys., vol. 4, p1048, 1978.

[166] A. Philipossian and D. Jackson, J. Electrochem. Soc., vol. 139, p. L82, 1992.

[167] C.S. Rafferty et al., "Explanation of Reverse Short Channel Effect by Defect Gradients", IEEE IEDM, p.311, 1993.

[168] A. T. Wu et al., "Nitridation Induced Surface Donor Layer in Silicon and its impact on the Characteristics of n- and p- channel MOSFETs", IEEE IEDM, p.271, 1989.

[169] J.R. Brews, "Subthreshold Behavior of Uniformly and Nonuniformly Doped Long-Channel MOSFET", IEEE Trans. Elect. Dev., ED-26, p.1282, 1979.

[170] C. Subramanian et al., "Reverse Short Channel Effect and Channel Length Dependence of Boron Penetration in PMOSFETs", IEEE IEDM, p.423, 1995.

[171] A.B. Joshi et al., "Suppressed Process-Induced Damage in N_2O-Annealed SiO_2 Gate Dielectrics", IEEE IRPS, p.156, 1995.

[172] K.Mistry, Private Communications.

[173] T. Hori and H. Iwasaki, "Ultra-thin Re-oxidized Nitrided Oxides Prepared by Rapid Thermal Processing", IEEE IEDM, p.570, 1987.

[174] T.S. Sriram and J. Bedard, "Correlation of RMS microroughness to TZDB measurements for oxides grown on nitrogen implanted Si", DEC Internal Memo, May 24th 1995.

[175] T.S. Sriram and J. Bedard, "AFM characterization of Si microroughness evolution with processing", DEC Internal Memo, Feb. 27th 1995.

[176] T. Ohmi et al. "Dependence of Thin-Oxide Films Quality on Surface Microroughness", IEEE Trans. Elect. Dev., ED-39, p.537, 1992.

[177] B.J. Fishbein, "Measurement of CMOS4 p-channel gate insulator reliability on lot A40048", DEC Internal Memo, Feb. 15th, 1991.

APPENDIX

"Characterization of Silicon Oxidation Under Low Dose N2 Implantation for Ultra-Thin Gate Oxides" by M.Rennie, H.R. Soleimani and B.S. Doyle, Accepted for publication in the Journal of Electrochemical Soc. 1996.

"Impact of Interface Roughness on Channel Mobility and Dielectric Breakdown with Gate Oxides Grown on N2-Implanted Silicon" by M. Rennie, H.R. Soleimani, B.S. Doyle and T.S. Sriram, Submitted to the IEEE Electron Device Letters for publication.

"Gate Oxide Control by Pre-Oxide Implantation of Nitrogen" by M. Rennie, H.R. Soleimani and B.S. Doyle, Submitted to the IEEE Proc. IEDM 1996.

Characterization of Silicon Oxidation Under Low Dose N_2 Implantation for Ultra-Thin Gate Oxides

M. Rennie, H.R. Soleimani*, and B.S. Doyle**

Digital Semiconductor, Reliability Engineering Group, South Queensferry, Scotland, UK.

**** Digital Equipment Corporation, ULSI Operations Group, Hudson, MA, 01749 USA***

***** Now at Intel, 2200 Mission College Blvd, Santa Clara, CA, USA.***

Aabstract

Silicon oxidation, in the presence of nitrogen atoms in Si, has been studied for very thin oxides. Prior to oxidation, nitrogen atoms have been incorporated into the substrate by N_2 implantation through a 225 Å thick screen/sacrificial oxide. After removing the screen oxide, the implanted wafers have undergone dryO_2 oxidation. Well behaved oxidation characteristics have been obtained for 10, 20 or 30 keV implant energies, and low 10^{14} - 10^{15} cm^{-2} N_2 doses. The results show a systematic reduction in the oxidation rate with nitrogen dose or energy, which in support of earlier work indicates that retardation in the oxidation rate is directly related to the nitrogen content at the Si/SiO_2 interface. The fact that the growth rate is retarded by the nitrogen atoms coming from the Si substrate and not supplied by an ambient gas such as N_2O from the top surface, proves that the retardation is a surface reaction and not a diffusion limited phenomenon.

Introduction

Direct implantation of nitrogen into silicon (Si) has been recently proposed [1] as a nitridation technique for the fabrication of ultra-thin, hot carrier--resistant SiO_2 gate oxides. The technique provides a manufacturable process for incorporating the desired amount of N atoms at the Si/SiO_2 interface, and it can easily be integrated into an existing manufacturing process. In that work [1], Si oxidation in the presence of nitrogen atoms in Si, was studied for a process that included the following steps:

1. implantation of N_2 into Si through a sacrifical SiO_2 oxide

2. a high temperature annealing (HTA) step prior to gate oxidation

3. removal of the sacrificial oxide

4. the actual gate oxidation

The pre-gate HTA step (step 2) was used in order to pile up the nitrogen atoms at the interface for the beginning of the gate oxidation step, and also anneal out any ion implantation-induced damage.

In the same wrok [1], another group of samples were subjected to the aforementioned processing steps with the exception of the step 2, which resulted in grearter retardation of Si oxidation (indicating a greater nitrogen pile-up at the interface). This, however, was only a feasibility study and did not investigate oxidation under various processing conditions such as different ion implantation dose/energy and/or oxidation time.

The aim of the present work is to characterize Si oxidation in connection with the direct-N_2-implantation nitridation technique, without the use of a pre-gate HTA step. The advantage of this approch is to eliminate one processing step (i.e. the pre-gate anneal), reduce the etch time of the sacrificial oxide[1], and with the use of a lower N_2 dose to avoid possible high dose implantation-induced damage.

Experiments

The silicon substrates used in this study were 150 mm, p-type <100> CZ wafers with resistivities ranging from 20 to 40 Ω-cm. They had an apparent native oxide thickness of 15 Å (ellipsometric analysis) which corresponded to an actual native oxide thickness of 6-8 Å. The wafers were RCA cleaned, and a 225 Å dryO_2 screen oxide was grown on the substrates prior to N_2 implantations. Three different implant energies at various N_2 doses were investigated as detgailed in Table 1. All implantations were performed at a 7^o tilt angle. Following N_2 implantation the screen oxide was removed by etching in 1:10 buffered HF acid for 10 minutes. The samples were subsequently oxidized at atmospheric pressure for 3, 5, 8, and 12 minutes in dryO_2 at 900 ^{o}C for each of the implant conditions in Table 1. After oxidation a 30 min anneal at 900 ^{o}C in N_2 was used to reduce the oxide fixed charge. A pure N_2 ambient was used to load and unload the wafers at 800 ^{o}C. Ramp-up's and ramp-down's were also performed in a pure N_2 ambient.

[1] After a pre-gate HTA step the sacrificail oxide etch rate reduces significantly due to the piled up nitrogen atoms [1,2].

All anneals and oxidations performed in this study were carried out in a commercially available atmospheric thermal silicon oxidation system. Dielectric thickness was measured at nine points on each wafer using standard ellipsometry. All of the results reported here represent the 9-point average of the measured values. The intra-wafer thickness uniformities were within ± 3 Å for doses less than 2 X 10^{15} cm^{-2}, and around ±8 Å at the highest dose. Although there might be some changes in dielectric constant due to the interfacial nitrogen (which might in turn affect the accuracy of the thickness readings), a previous study of RNO high nitrogen content oxides has shown that there is no noticeable difference between ellipsometer readings and TEM micrograph measurements [3].

Results and Discussions

The concentration of nitrogen at the interface at the beginning and during the oxidation process is extremely important in controlling the oxidation kinetics. Nitrogen atoms incorported into Si by ion implantation, have been reported [4] to evaporate from bare Si substrates under high temperature non-oxidizing annealing conditions. But, in the presence of a SiO_2 oxide N atoms pile up at the Si/SiO_2 interface and do not evaporate. To verfiy the formation of a pile-up and the lack of N evaporation, N_2 atoms were implanted through a 450 Å thick SiO_2 oxide and were subsequently annealed at 975 °C for different times. The SIMS profiles (Fig.1) show the as-implanted and annealed nitrogen profiles from these samples. The annealed profiles exhibit both the formation of a pile-up at the interface, and the neglegible N diffusion through the 540 Å oxide.

Figures 2-4 show the oxide thickness versus N_2 dose for three different energies. For 10 keV N_2 implant energy (Figure 2) it can be seen that a dependence of oxide growth on N_2 dose begins at around 4 x 10^{14} cm^{-2} dose below which the oxide thickness (T_{ox}) is practically the same as the corresponding control oxide T_{ox}. At the higher N_2 implant energies (Figures 3 & 4) retardation in the oxide growth is evident to happen at even smaller doses (note that the oxide thicknesses at 2 x 10^{14} cm^{-2} dose in Figure 3 are almost the same as the oxide thicknesses at 1 x 10^{15} cm^{-2} dose in Figure 2.) In Figures 3 & 4, at doses higher than 1 x 10^{15} cm^{-2}, oxide growth is almost non-existant. This behavior can be more clearly seen in Figures 5 & 6 which show the T_{ox} as a function of oxidation time for different N_2 doses (Figures 5 & 6 correpond to Figures 2 & 3, respectively).

Comparing Figure 2 to Figures 3 & 4, it can be seen that at 10 keV it takes a much higher dose to retard the oxidation compared to the 20 & 30 keV implants. The significantly smaller retardation in the oxide growth rate for the 10 keV implanted samples, is due to the fact that the peak of the N_2 implant is located within the 225 Å screen oxide, and as a result a major portion of N_2 dose is lost when the screen oxide is subsequently stripped off.

The fact that the oxide growth rate is retarded by the nitrogen atoms coming from the Si substrate and not supplied by an ambient gas such as N_2O from the top surface, proves that the retardation is a surface reaction and not a diffusion limited phenomenon.

The fact that the oxide growth rate is retarded by the nitrogen atoms coming from the Si substrate and not supplied by an ambient gas such as N_2O from the top surface, proves that the retardation is a surface reaction and not a diffusion limited phenomenon.

Conclusion

Si oxidation, in the presence of nirogen atoms in Si, is studied for hot carrier-resistant gate oxide and mixed voltage circuit [6] applications. Nitrogen atoms are incorporated into the Si substrate by N_2 implantation prior to the gate oxidation step. Oxidations under different N_2 energies or doses were examined for four oxidation times, and they were compared against corresponding dryO_2 oxidations. The results show a systematic reduction in the oxidation rate with nitrogen dose or energy, which in support of earlier work indicates that retardation in the oxidation rate is directly related to nitrogen. The fact that the growth rate is retarded by the nitrogen atoms coming from the Si substrate and not supplied by an ambient gas such as N_2O from the top surface, proves that the retardation is a surface reaction and not a diffusion limited phenomenon.

Acknowledgments

The authors would like to thank Ian Maitland for the nitrogen implantations. SIMS analyses were done by Evans Europa.

Reference

[1] H.R. Soleimani, B.S. Doyle, and A. Philipossian, Electrochem Society Lett., submitted.

[2] G. Weinder, R. Kurps, K. Blum, and D. Kruger, *infos'93*, Elsevier, p. 77, 1993.

[3] A. Philipossian and D. Jackson, J. Electrochem Soc., vol. 139, p. L82, 1992.

[4] W.J.M.J. Josquin, Nuclear Instruments and Methods, 209/210, p. 581, 1983.

[5] M. Ramin, H. Ryssel and H. Kranz, Apply. Phys. 22, p. 393, 1980.

[6] M. Rennie, B.S. Doyle and H.R. Soleimani, to be submitted to IEDM 1995.

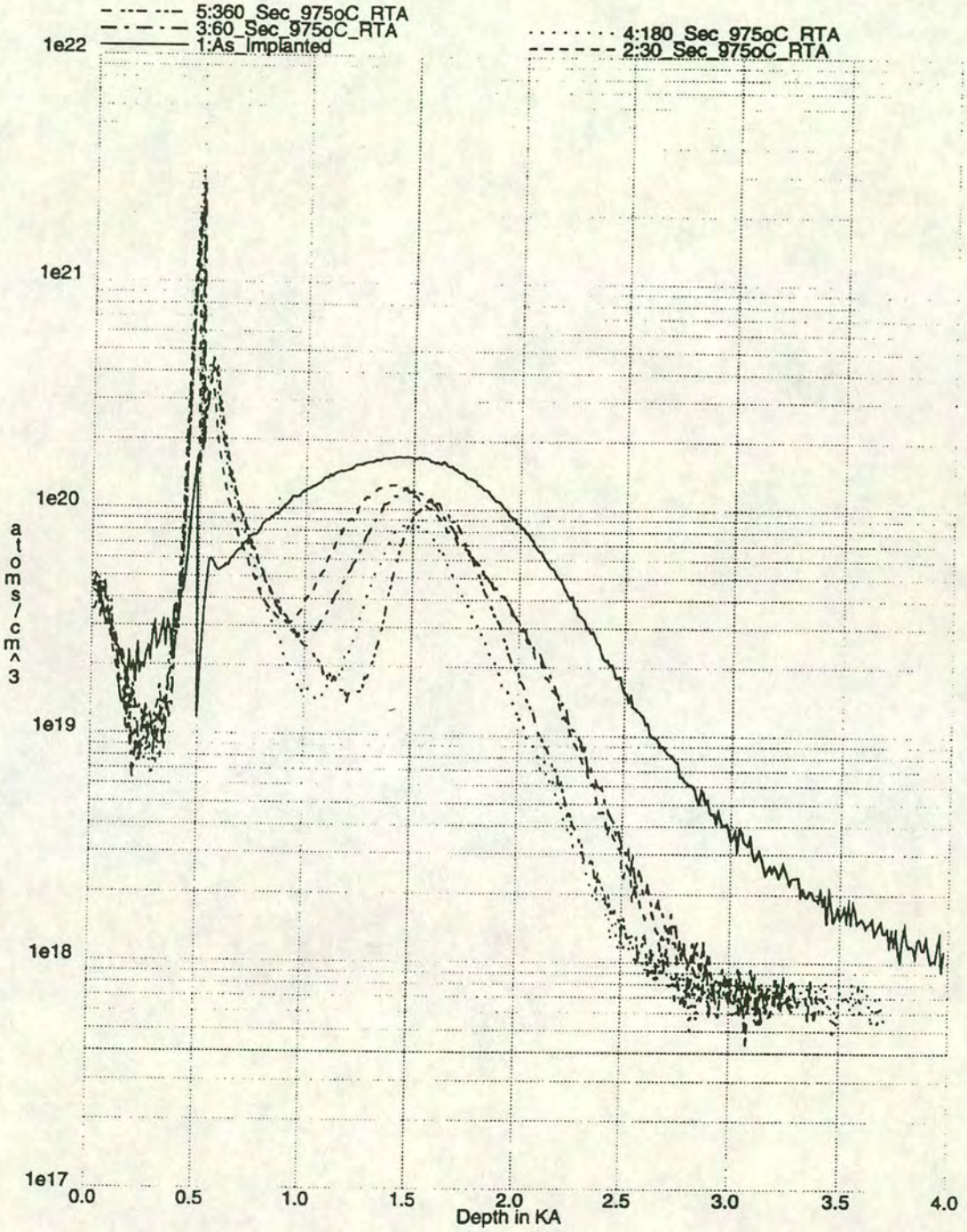


Figure 41 SIMS profiles of 10^{15} N_2^+ atoms/cm^2 ion implant at 100KeV in silicon through a 540Å screen oxide for (a) as-implanted, (b) 30 second RTA, (c) 60 second RTA, (d) 180 second RTA and (e) 360 second RTA at 975° C.

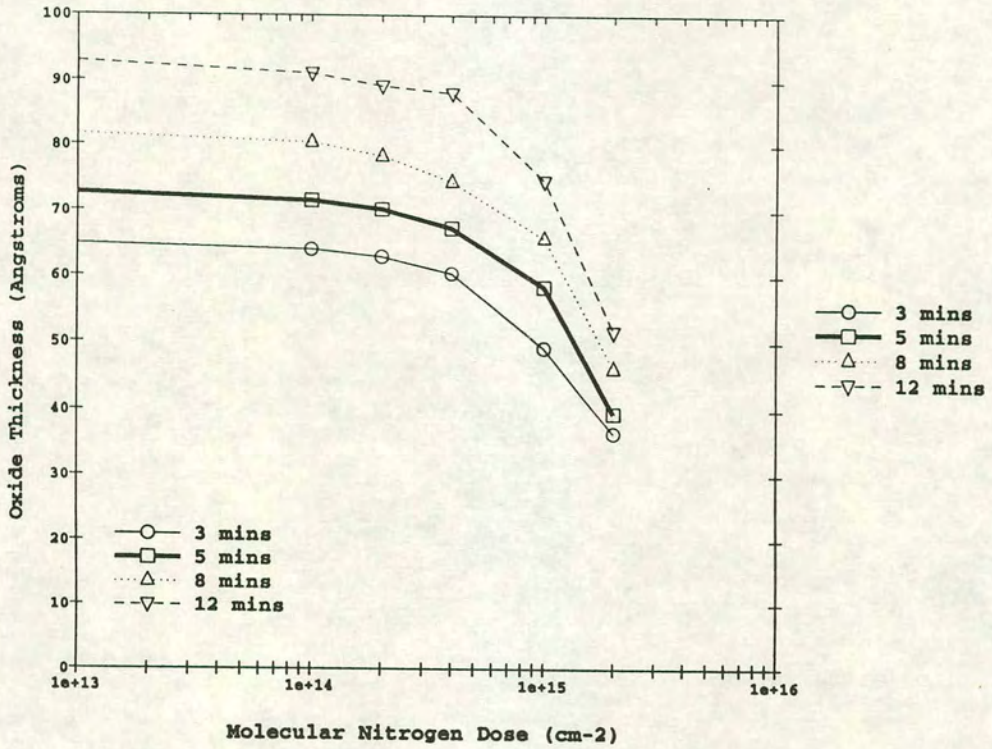

Oxide Thickness (Angstroms)
100
90
80
70
60
50
40
30
20
10
0
1e+13
1e+14
1e+15
1e+16
Molecular Nitrogen Dose (cm-2)
3 mins
5 mins
8 mins
12 mins
3 mins
5 mins
8 mins
12 mins

FIG 2

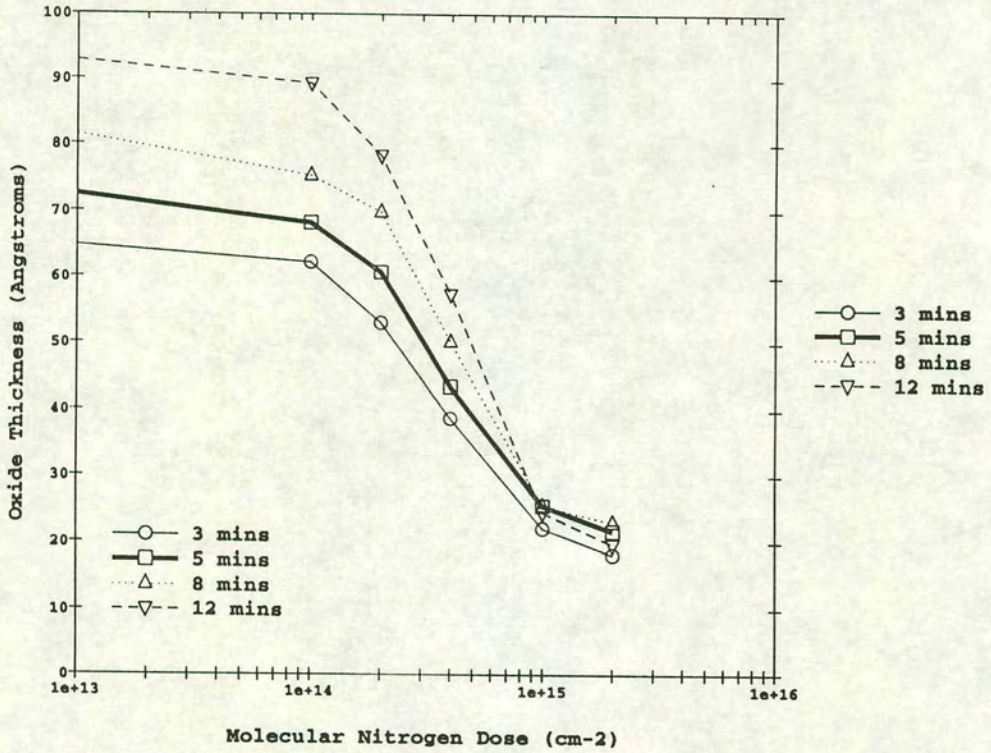
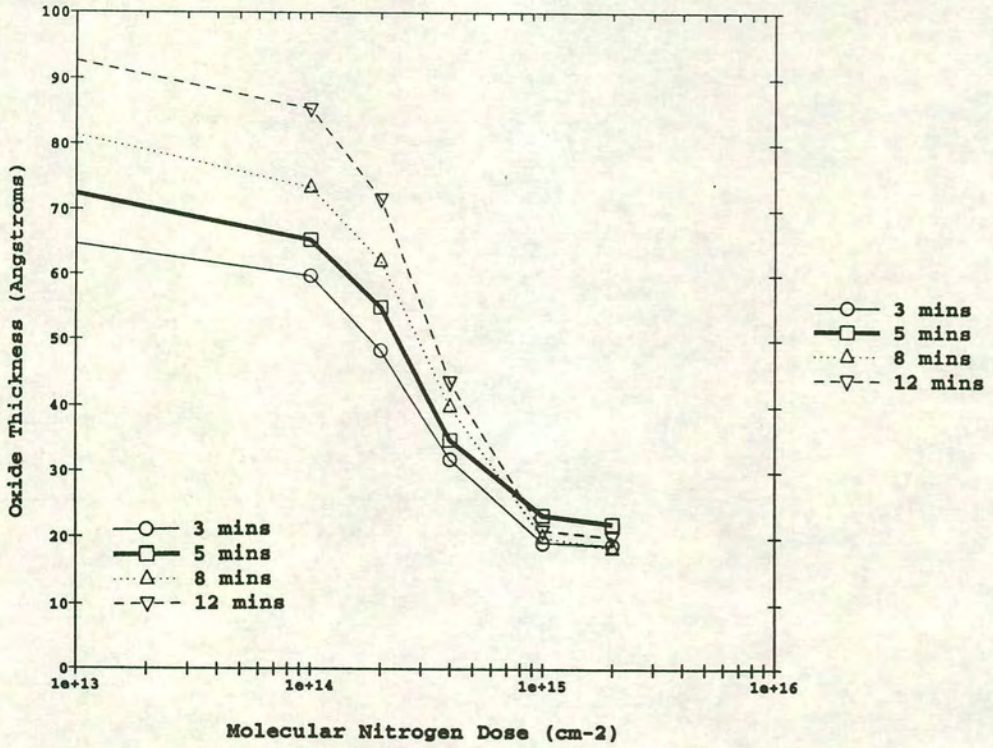

Oxide Thickness (Angstroms)
100
90
80
70
60
50
40
30
20
10
0
1e+13
1e+14
1e+15
1e+16
Molecular Nitrogen Dose (cm-2)
3 mins
5 mins
8 mins
12 mins
3 mins
5 mins
8 mins
12 mins

Figure 4


Oxide Thickness (Angstroms)
100
90
80
70
60
50
40
30
20
10
0
1e+13
1e+14
1e+15
1e+16
Molecular Nitrogen Dose (cm-2)
3 mins
5 mins
8 mins
12 mins
3 mins
5 mins
8 mins
12 mins

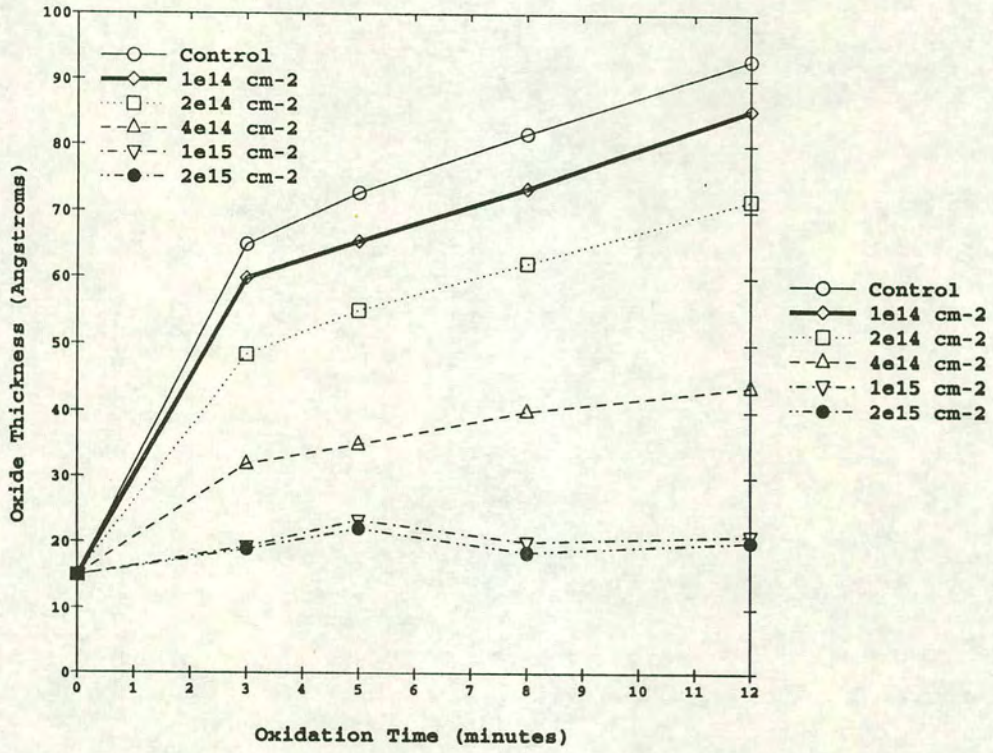

Control
1e14 cm-2
2e14 cm-2
4e14 cm-2
1e15 cm-2
2e15 cm-2
Oxide Thickness (Angstroms)
Oxidation Time (minutes)
0
10
20
30
40
50
60
70
80
90
100
0
1
2
3
4
5
6
7
8
9
10
11
12

Figure 5

Figure 2

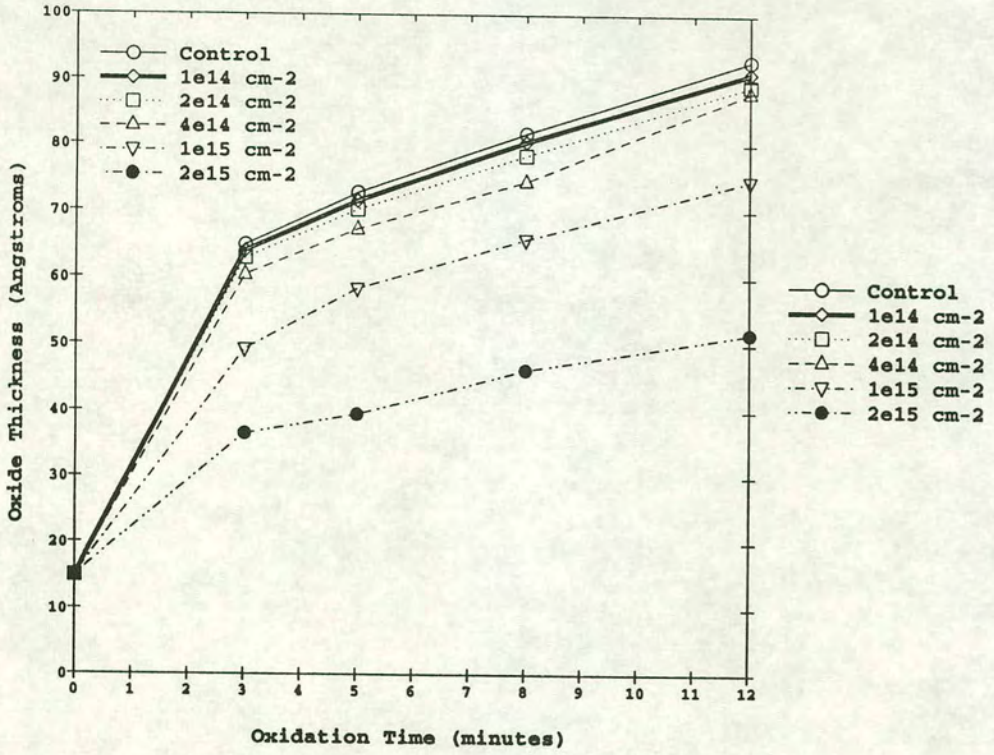

Control
1e14 cm-2
2e14 cm-2
4e14 cm-2
1e15 cm-2
2e15 cm-2
Oxide Thickness (Angstroms)
Oxidation Time (minutes)
0
10
20
30
40
50
60
70
80
90
100
0
1
2
3
4
5
6
7
8
9
10
11
12

Figure 7

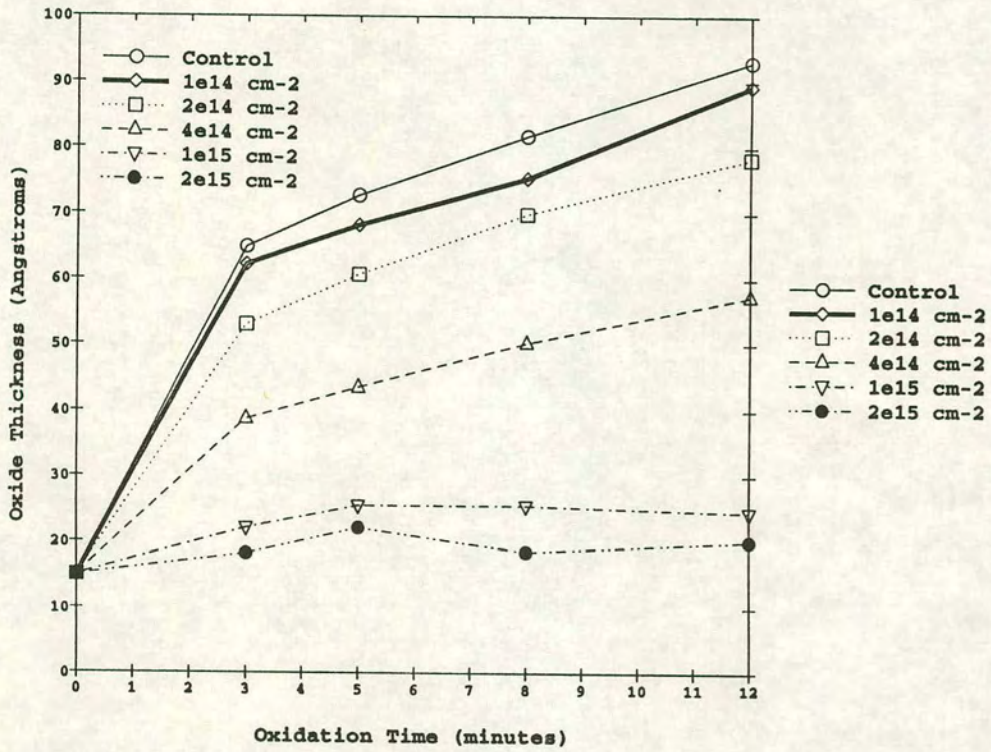
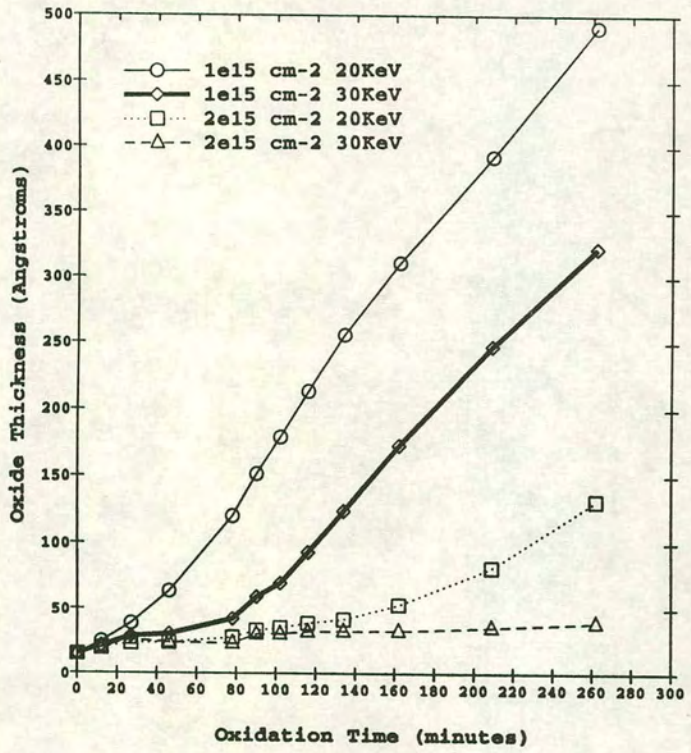

Control
1e14 cm-2
2e14 cm-2
4e14 cm-2
1e15 cm-2
2e15 cm-2
Oxide Thickness (Angstroms)
Oxidation Time (minutes)
0
10
20
30
40
50
60
70
80
90
100
0
1
2
3
4
5
6
7
8
9
10
11
12

FIGURE 8



1e15 cm-2 20KeV
1e15 cm-2 30KeV
2e15 cm-2 20KeV
2e15 cm-2 30KeV

Impact of Interface Roughness on Channel Mobility and Dielectric Breakdown with Gate Oxides Grown on N_2-Implanted Silicon.

M. Rennie†, H.R. Soleimani, B.S. Doyle‡, and T.S. Sriram

Digital Equipment Corporation, ULSI Operations Group, Hudson, MA, 01749 USA

† Was with Digital Semiconductor, South Queensferry, Scotland, U.K.
Now with Analog Devices, Wilmington, MA 01887, U.S.A.

‡ Now with Intel Corp. in Santa Clara, CA, USA.

May 23, 1995

Abstract

In this paper we show the strong correlation of silicon/silicon oxide interface microroughness with MOSFET channel mobility and oxide breakdown field. The gate oxide samples used in this study were formed by the oxidation of nitrogen implanted silicon. We present cross-sectional TEM evidence to show that the increased interface roughness is due to localized thinning of the gate oxide and proposed the mechanism is the inhomogeneous redistribution of nitrogen atoms.

Introduction

Microroughness and defects at the Si-SiO_2 interface have a large affect on the breakdown characteristics of thin <100> oxides [1,2]. These imperfections can also cause loss of MOSFET drive current through reduced carrier mobility [3]. Previously, much work has concentrated on the introduction of nitrogen into gate oxides by oxidation in the presence of a nitrogen containing ambient such as NH_3[4], N_2O [5] or NO [6]. Alternatively Kuroi *et al.* [7] have shown improvements to device performance and reliability with the implantation of nitrogen through the defined poly gate. Recently [8], we have characterized the growth kinetics of low dose nitrogen implanted silicon. In that study, we provided evidence that nitrogen evaporates from the silicon surface in an oxidizing ambient until the concentration is low enough to allow oxidation to proceed. Once an oxide layer is formed on the surface, the nitrogen piles up to the Si-SiO_2 interface due to the lower nitrogen diffusitivity through the oxide. Using this technique, it has been proposed that a greater efficiency of nitrogen incorporation at the Si-SiO_2 interface can be achieved than the aforementioned techniques [9]. The purpose of this study is to determine the conformality of grown gate oxide and resulting oxide properties with incorporating the nitrogen by direct ion implantation into the silicon substrate.

Experiments and Results

MOSFET devices used in this work were fabricated using a baseline 0.5 μm CMOS process. Nitrogen ions (N_2^+) were implanted through a 150 Å sacrificial oxide after threshold implants and shallow trench isolation. A standard RCA clean was used to etch the screen oxide and conventional furnace oxidations were carried out at 900 °C in a dry O_2 ambient at atmospheric pressure to grow 100 Å gate oxides. The wafers were pulled in and out of the furnace in a pure nitrogen ambient and a 60 min 900 °C anneal in nitrogen was used to reduce the

fixed oxide charge. As previously characterized [8], the oxidation times were adjusted according to nitrogen dose. All wafers were then processed in the same batch through transistor formation and 4 level metal interconnect. Samples required for Atomic Force Microscopy (AFM) were processed with the device wafers through sacrificial oxidation, nitrogen implantation, RCA clean and gate oxide growth. AFM measurements were conducted on the oxide surface and on the silicon surface after the gate oxide was etched off. Table 1 shows the nitrogen implant split conditions and AFM RMS results before and after gate oxide strip. A clear trend of increasing microroughness with nitrogen implant conditions is observed. The gate oxide surface is also measured to be rougher than the silicon substrate surface. To determine the affect on channel mobility, both nMOSFET and pMOSFET devices were measured as shown in Figures 1(a) and 1(b) respectively. Consistent with previous literature, the channel high field mobility degrades as the microroughness increases, we do not understand at this point however why the low field mobility is also affected. Voltage ramp dielectric breakdown measurements were taken by applying 0.1 second gate voltage pulses in steps of 0.5 MV/cm. Alternating low field (3.5 MV/cm) measurement pulses were used to determine dielectric failure at 1mA leakage current limit. nMOSFET and pMOSFET de vices with a polysilicon area antenna ratio of 1, were stressed in accumulation mode. Figures 2(a) and 2(b) show lower field breakdown distributions with increasing microroughness. These results suggest a non-uniform thickness of gate oxide is grown where the non-uniformity increases with increasing nitrogen dose. In order to determine if the increasing microroughness, degrading mobility and reduced breakdown field is attributable to unannealed ion implant damage or increasing nitrogen content, cross-sectional Transmission Electron Microscopy (XTEM) was performed on sample F. As seen in Figure 3, there is a localized thinning of the grown oxide which we propose is due to the non-uniform redistribution of N atoms during the gate oxidation. Ion implantation damage to the silicon would not account for the non-conformal oxide and periphery intensive diode junction breakdown measurements were similar for samples A to F. Previous measurements of oxide breakdown on oxides grown directly in a N_2O am-

bient gave catastrophic results compared to nitridation after the growth of an conventional oxide layer [10]. The mechanism suggested in that work was the imbalance between growth and volatilization in the N_2O ambient which resulted in islanded growth from Si-N bonding, we conclude that the results of this work support this mechanism albiet the nitrogen was incorporated by ion implantation.

Conclusion

We have shown the strong correlation of gate oxide breakdown field, channel mobility and Si-SiO_2 interface roughness with the use of oxidized nitrogen implanted silicon. An increased nitrogen dose is seen to roughen the interface due to the localized thinning of the as-grown oxide. We propose the non-uniform oxide thickness results from the inhomogeneous redistribution of N atoms. Oxides grown with relatively low dose nitrogen ($\approx 10^{14}$ cm^{-2}) implanted silicon have device performance and reliability properties comparable to conventional oxide.

Reference

[1] T. Ohmi, M. Miyashita, M. Itano, T. Imaoka and I. Kawanbe, *Dependance of Thin-Oxide Films Quality on Surface Microroughness*, IEEE Trans. Elect. Dev., Vol. 39, No. 3, p537, 1992.

[2] P.O. Harn, M. Grandner, A. Schnegg and H. Jacob, in The Physics and Chemistry of SiO_2 interface, edited by C.R. Helms and B.E. Deal, Plenum, New York, p401., 1988.

[3] T. Ohmi, K. Kotani, A. Termoto and M. Miyashita, *Dependance of electron mobility on Si-SiO_2 interface microroughness*, IEEE Elec. Dev. Lett., Vol. 12, No. 12, p. 652, 1991.

[4] T Hori and H. Iwasaki, *Improved Hot-Carrier Immunity in Submicrometer MOSFETs with Reoxidized Nitrided Oxides Prepared by Rapid Thermal Processing*, IEEE Elec. Dev. Lett., Vol 10, No. 2, p. 64, 1989.

[5] H.R. Soleimani, A. Philipossian and B.S. Doyle, *A Study of the Growth Kinetics of SiO_2 in N_2O*, IEDM Extended Abstracts, p. 402, 1992.

[6] Y. Okada, P.J. Tobin, K.G. Reid, R.I. Hedge, B. Maiti and S.A. Ajuria, *Gate Oxynitride Grown in Nitric Oxide (NO)*, Tech. Dig. Symp. VLSI Tech., p. 105, 1994.

[7] T. Kuroi, T. Yamaguchi, M. Shirahata, Y. Okumura, Y. Kawasaki and N. Tsubouchi, *Novel NICE (Nitrogen Implantation into CMOS Gate Electrode and Source-Drain) Structure for High Reliability and High Performance 0.25 μm Dual Gate CMOS*, IEDM Extended Abstracts, p. 325, 1993.

[8] M. Rennie, H.R. Soleimani and B.S. Doyle, *Characterization of Silicon Oxidation under Low Dose N_2 Implantation for Ultra-Thin Gate Oxides*, J. Electrochemical Society, submitted.

[9] H.R. Soleimani, B.S. Doyle and A. Philipossian, *Formation of Ultra-Thin Nitrided SiO_2 Oxides by Direct Nitrogen Implantation Into Silicon*, J. Electrochemical Society, submitted.

[10] A. Philipossian and B.S. Doyle, *The Effect of Pre-Oxidation Ambient on the Electrical Integrity of N_2O Grown Gate Dielectrics*, unpublished results.

Sample ID	N_2 Implant Conditions	Gate Oxidation Time (min)	Gate Oxide Surface RMS roughness (nm)	Silicon Surface RMS RMS roughness (nm)
A	No Implant (Conventional Oxide)	12	0.165-0.176	0.153-0.159
B	1.0e14 atoms/cm2 10KeV	14	0.192-0.198	0.178-0.185
C	1.2e14 atoms/cm2 10KeV	18	0.211-0.212	0.179-0.181
D	2.0e14 atoms/cm2 17KeV	28	0.276-0.282	0.237-0.269
E	3.0e14 atoms/cm2 25KeV	47	0.518-0.519	0.392-0.401
F	4.0e14 atoms/cm2 17KeV	53	0.530-0.519	0.415-0.433

Table 1: Nitrogen Implant Split conditions and Gate Oxidation times to grow 100 Å oxide. RMS roughness measurements are included for both AFM measurements on the Gate Oxide surface and on the Silicon surface after the oxide was etched off.

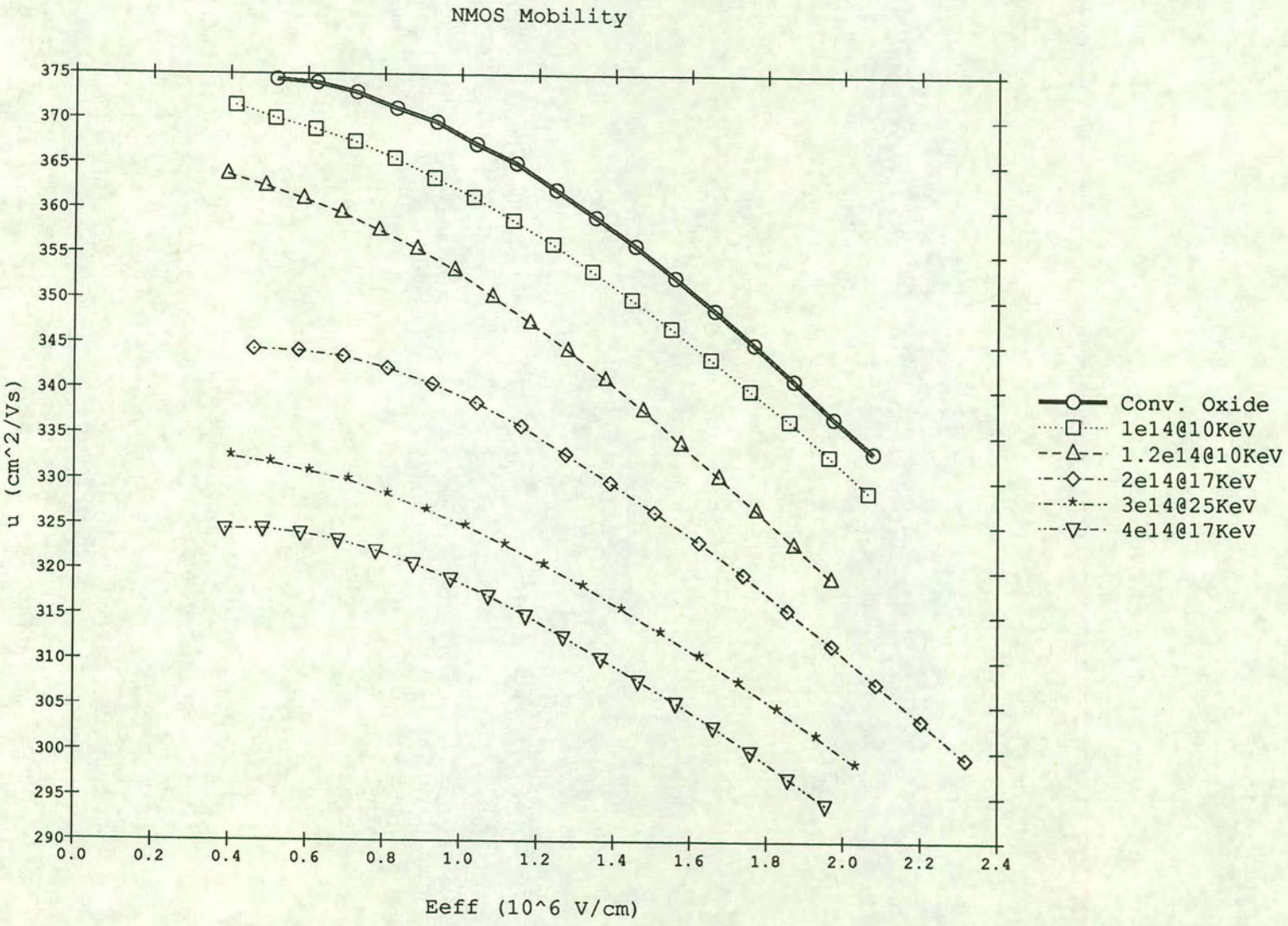

NMOS Mobility
u (cm^2/Vs)
Eeff (10^6 V/cm)
375
370
365
360
355
350
345
340
335
330
325
320
315
310
305
300
295
290
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
Conv. Oxide
1e14@10KeV
1.2e14@10KeV
2e14@17KeV
3e14@25KeV
4e14@17KeV

FIGURE 1(a)

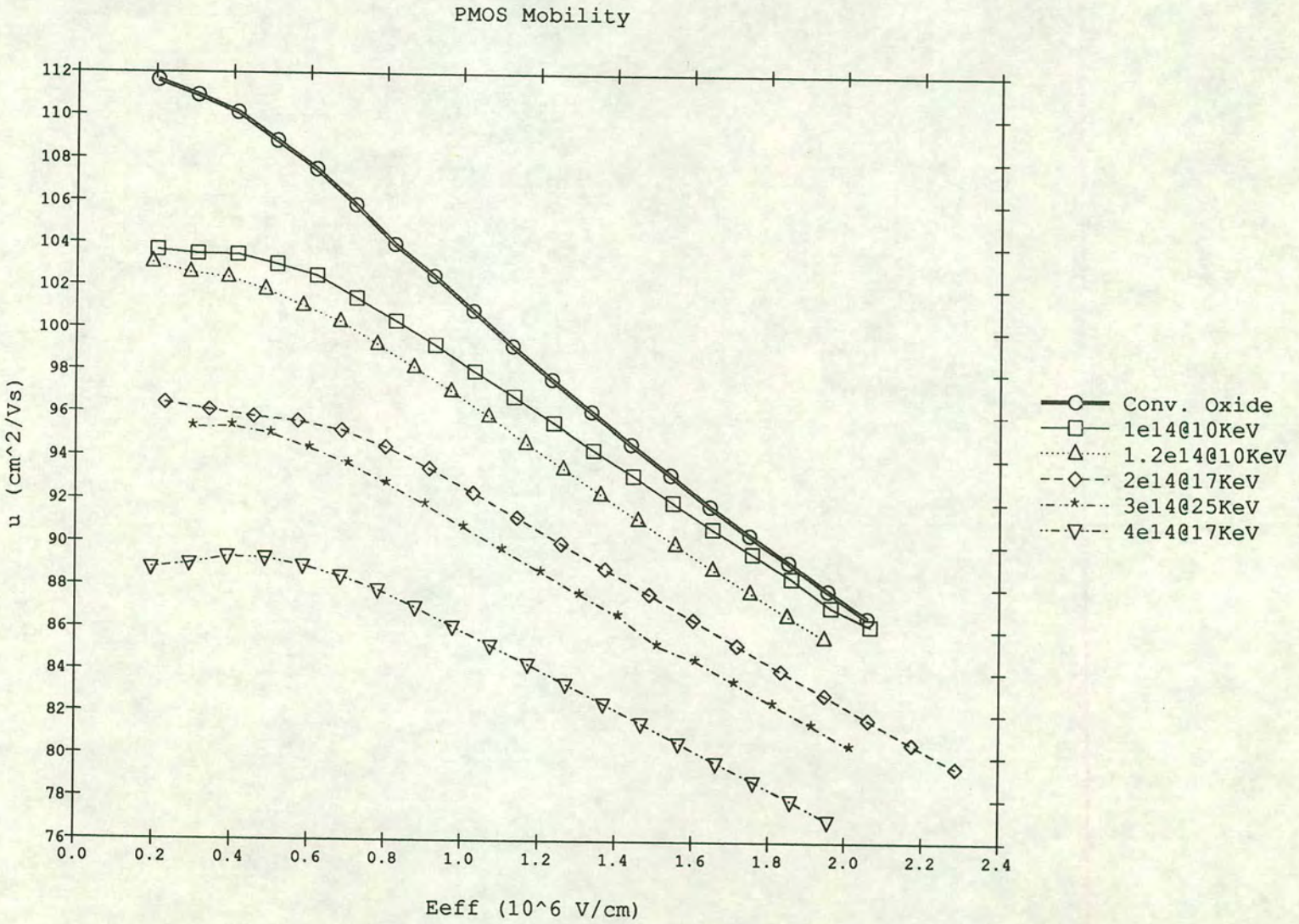

PMOS Mobility
u (cm^2/Vs)
Eeff (10^6 V/cm)
Conv. Oxide
1e14@10KeV
1.2e14@10KeV
2e14@17KeV
3e14@25KeV
4e14@17KeV

FIGURE 1 (b)

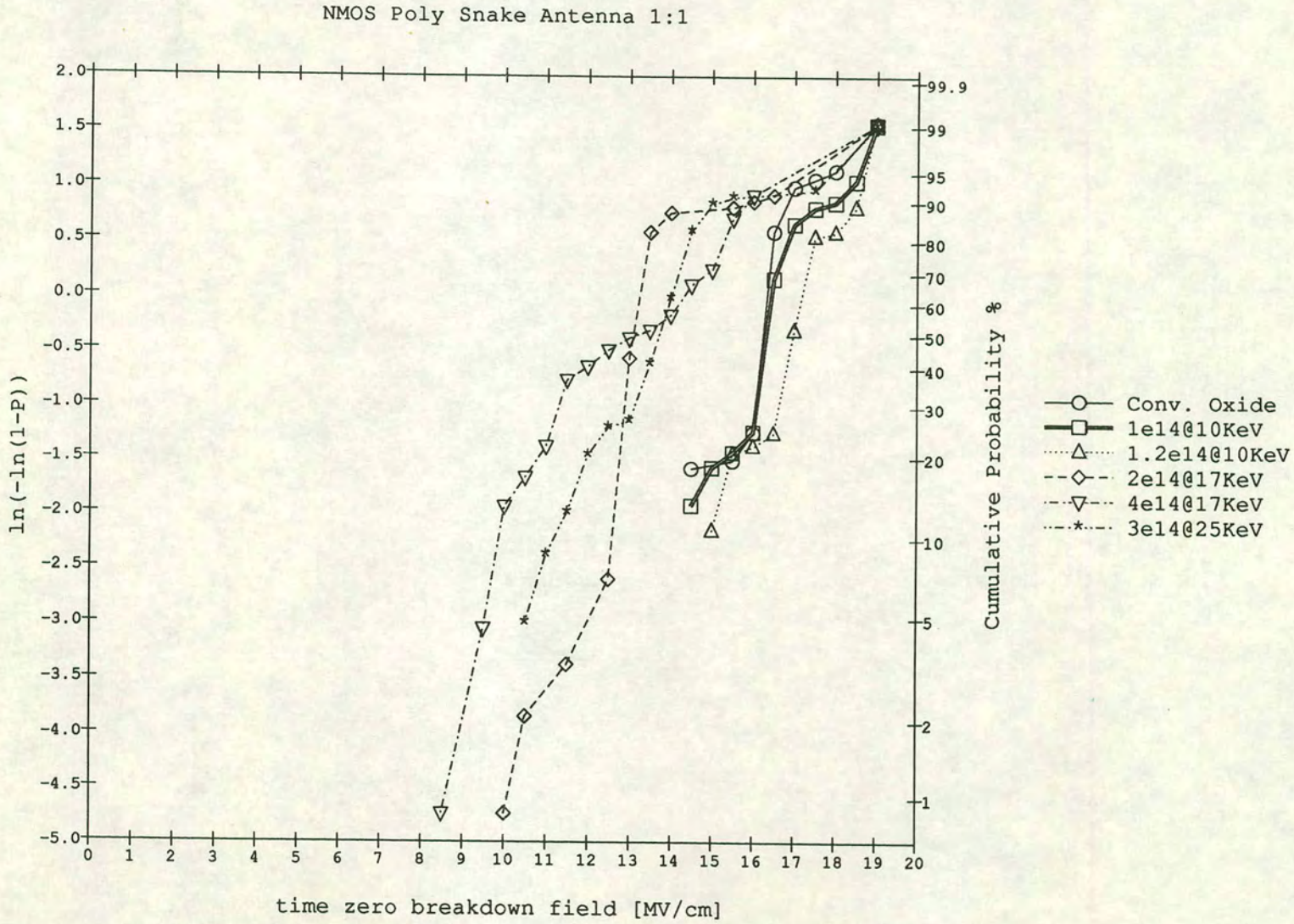
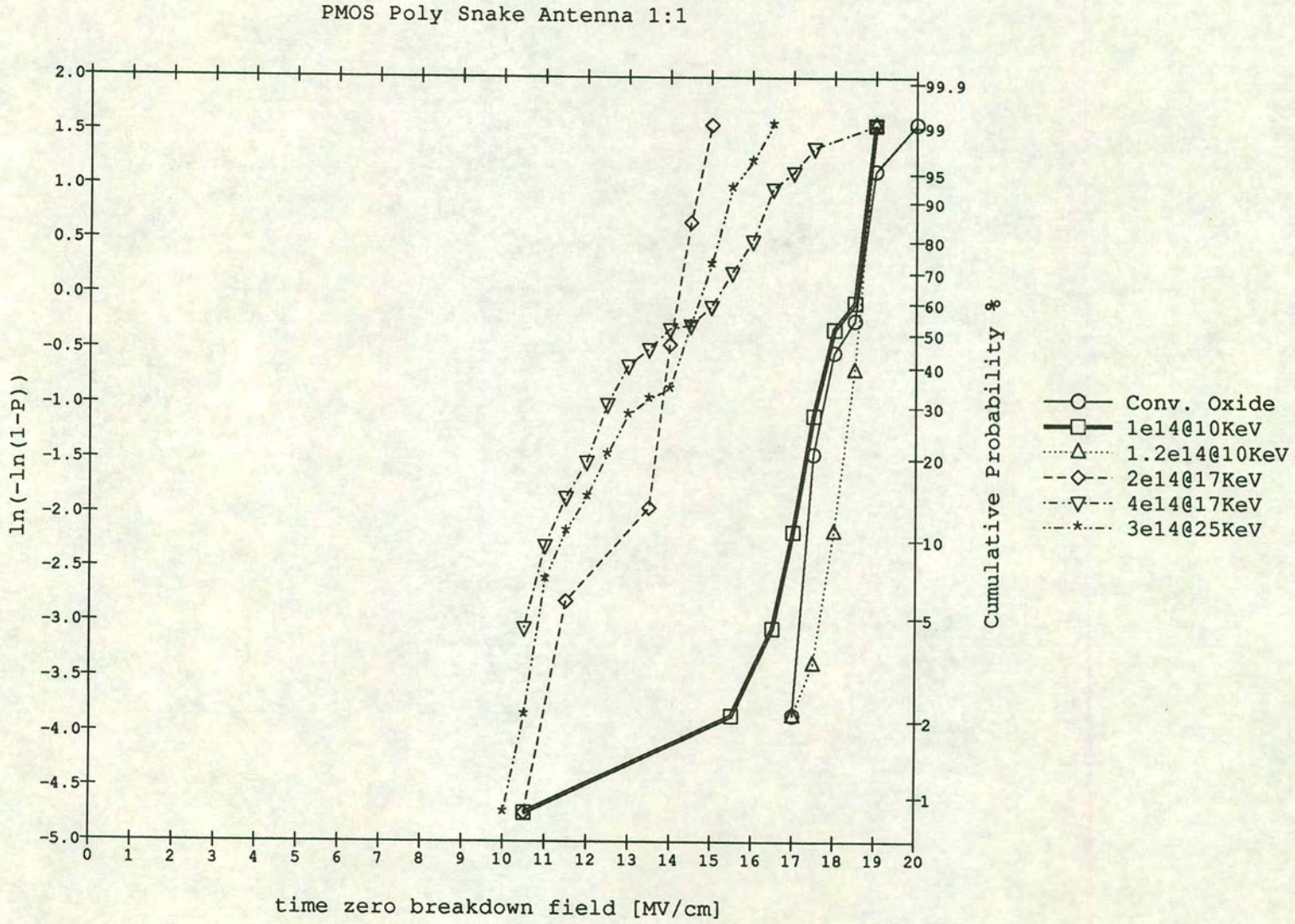
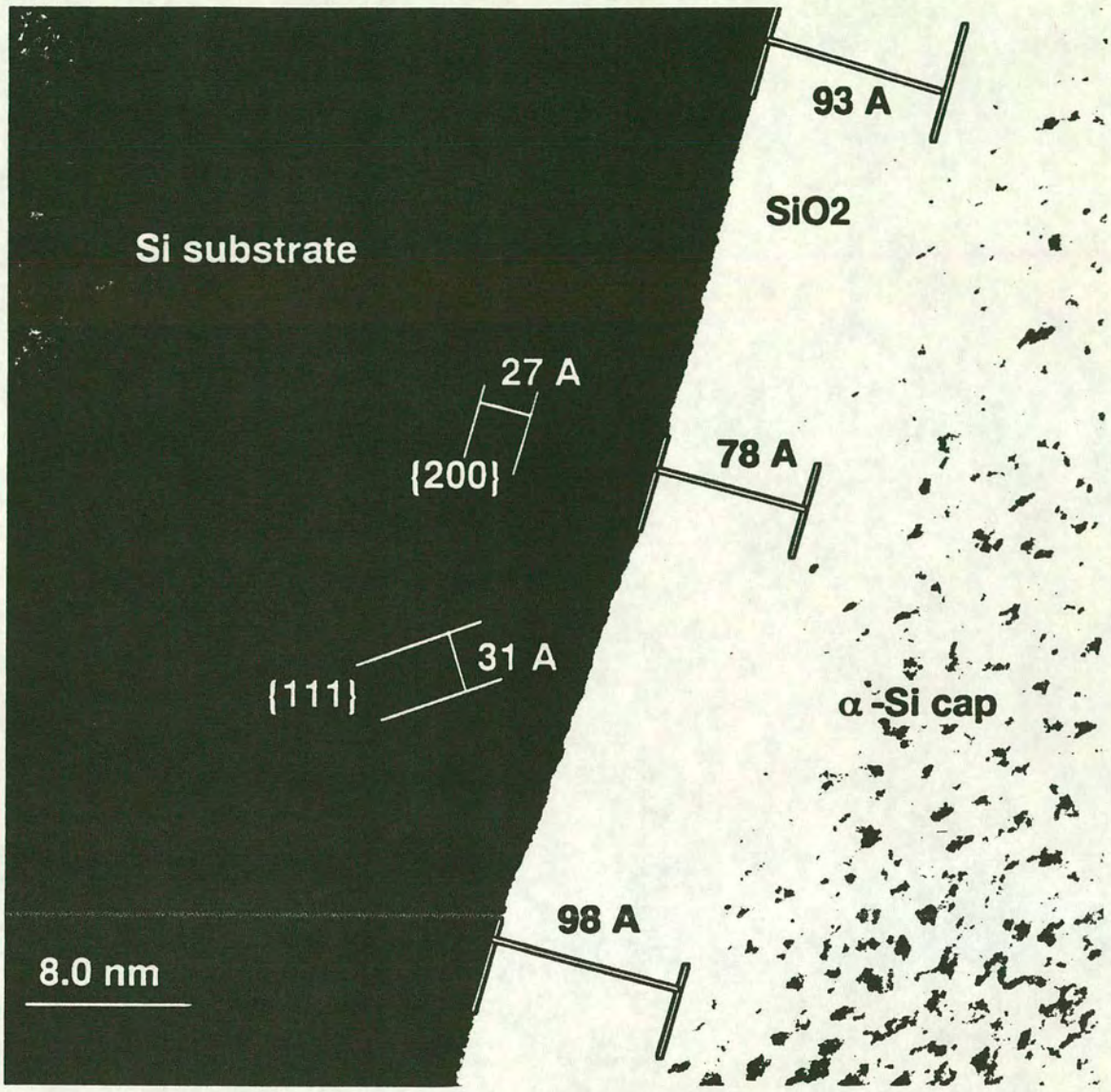

NMOS Poly Snake Antenna 1:1
ln(-ln(1-P))
Cumulative Probability %
time zero breakdown field [MV/cm]
Conv. Oxide
1e14@10KeV
1.2e14@10KeV
2e14@17KeV
4e14@17KeV
3e14@25KeV

FIGURE 7.(-)

Figure 7(b)


PMOS Poly Snake Antenna 1:1
ln(-ln(1-P))
Cumulative Probability %
time zero breakdown field [MV/cm]
Conv. Oxide
1e14@10KeV
1.2e14@10KeV
2e14@17KeV
4e14@17KeV
3e14@25KeV



High resolution cross-sectional TEM image from wafer 22 showing variations in oxide thickness.

FIGURE 3

Suggested Areas:- CMOS Devices & Reliability, Device and Interconnect Technology, Solid State Devices

Gate Oxide Control By Pre-Oxide Implantation of Nitrogen

by M.Rennie*, H.Soleimani and B.Doyle**

ital Equipment Corporation, 77 Reed Rd, Hudson, MA 01749, USA Tel:- (508)568-6443
w at Analog Devices, Wilmington, MA, USA, ** Now at Intel, Santa Clara, CA, USA

RODUCTION

ogen incorporation into the gate oxide has been used to give hot carrier hardness to these insulators for some time now]. One of the consequences of nitriding an oxide is that the growth rate of the formed oxide is significantly retarded [6]. paper explores a novel technique of nitriding the gate oxide to create oxides of different thicknesses on the same wafer. ontrast to other techniques, which nitride either during oxidation [3,4,6], or after oxidation[1,2,5], this novel technique oduces nitrogen into the silicon BEFORE oxidation, and involves the implantation of nitrogen into the silicon, which is done ctively in the same process step as Vt control implantations. It has applications in power/logic mixed mode, I/O drivers for c (to higher voltage peripherals), different logic/cache gate oxide thicknesses, etc, where two simultaneous gate oxide knesses are desirable. At high doses, the technique can also be used as an alternate method for producing hot carrier-l gate insulators.

TE OXIDE GROWTH

steps involved in the process are i) implantation of N2 through a sacrificial oxide (225 A - analogous to, and at the same as Vt implants) , ii) H.T. anneal prior to oxidation (optional), iii) removal of sacrificial oxide, and iv) gate oxidation itself. re 1 shows the SIMS data for 1E15, 100 keV implant followed by an anneal. As would be expected, there is uphill sion of nitrogen, and a piling up at the interface [8]. It is this nitrogen at the interface that slows down the oxidation rate, manner analogous to N2O oxidations [7]. The anneal step allows for the variation of the initial N dose at the interface, and centrations of up to 4E21/cm3or higher can be obtained [7]. A study of the implant conditions is given in Figures 2 and 3. n be seen that at the rate of oxidation can be approximated by a straight line, similar to N2O oxidations, and that the ter the dose the thinner the oxide, allowing the tailoring of gate oxide thicknesses. Figure 3 shows oxide thickness as a tion of dose. For doses up to 1E14/cm2, the thicknesses are identical to the non-implanted case. For large doses 5/cm2), the growth is effectively stopped. In the intermediate range, the oxide thickness can be effectively varied.

VICE CHARACTERISTICS

re 4 and 5 show typical characteristics for N-implanted devices. It can be seen that these devices behave like non N-anted transistors (the variation in characteristics is due to variations in Tox). Figure 6 & 7 show the mobility vs effective measurements. It can be seen that the higher the dose of implanted N, the lower the mobility. This lowering of the ility is seen throughout the field range. The reason for the low field lowering is not known (but is common to most nitriding niques), and could be related to the activation of nitrogen in the silicon itself (it is known that approx. 10% of N in silicon omes activated [8]). The reduction at high fields suggests that there is increased surface scattering, arising from greater ace roughness. AFM measurements of this interface confirm that this is so (Fig. 8).

LIABILITY

rougher surface from AFM might indicate that there might be some TDDB implications to this. However, neither the B, nor the Qbd measurements (included in Figures 9 and 10) show any significant difference between nitrogen implanted non-implanted oxides. The implants do not damage the junction edge in subsequent processing, as Figure 11 shows. The gen also acts as a barrier to boron penetration(Figure 12), even at these low concentrations of incorporated N. From the carrier data (Figures 13 and 14), it can be seen that for doses up to 1.2E14, there is no significant change in the hot er characteristics. This is not surprising since Figure 1 shows that the percentage of N incorporated is well below the 1% nally associated with increased hot carrier hardness. Preliminary results on the higher N incorporations show that a factor eater than 4 improvement is possible.

NCLUSIONS

stigation of a technique for the growth of gate oxides of different thicknesses through the implant of nitrogen into the n before gate oxide growth , in the same process step as Vt implants, are reported here. It is shown that this process s oxides that are as robust as conventional oxides, showing no decrease in gate oxide yield, hot carrier degradation nor ion leakage, and is thus attractive for applications that require dual gate oxide thicknesses.

FERENCES

/.Yang et al IEEE T.E.D. vol. ED-35, pp. 935-944, 1988.
.Hori et al, IEEE E.D.L. vol EDL-120, pp. 64-67, 1989
.Huang et al, IEDM Proc. pp. 421-424, 1990
.Tobin et al, Proc. VLSI Symp., p. 51, 1993
[5] T.Yamaguchi et al, IEDM Proc., p. 325, 1993
[6] H Soleimani et al. IEDM Proc., p. 629, 1992.
[7] H Soleimani et al. Accepted ECS Letters, Aug. 1995
[8] W.Josquin, Nucl. Inst. & Methd, vol. 209,p.581, 1984

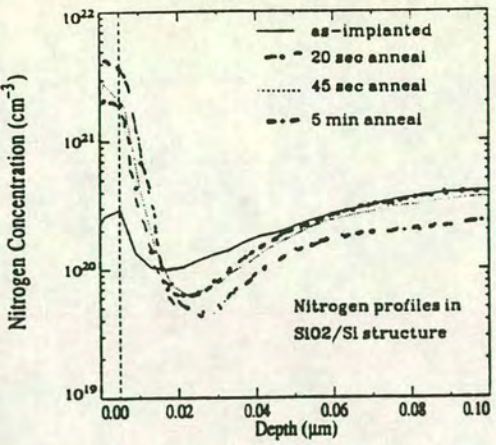


Figure 1. N profile at Si-SiO2 Interface as a function of anneal time (975 C)

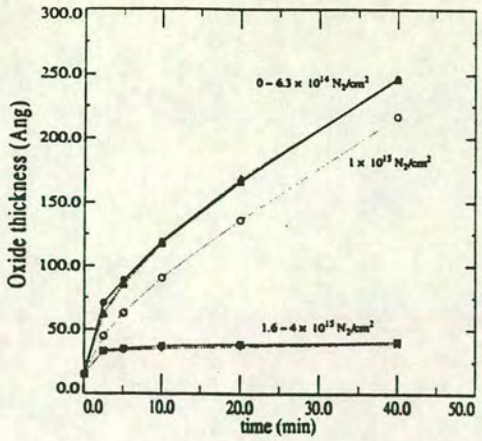


Figure 2. Oxide Thickness Vs Oxidation Time for Different N-Implant Doses

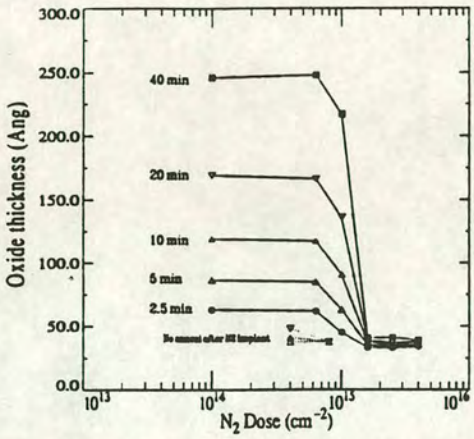


Figure 3. Data of Figure 2, plotted as Tox Vs N2 Dose, for different Oxidation Times

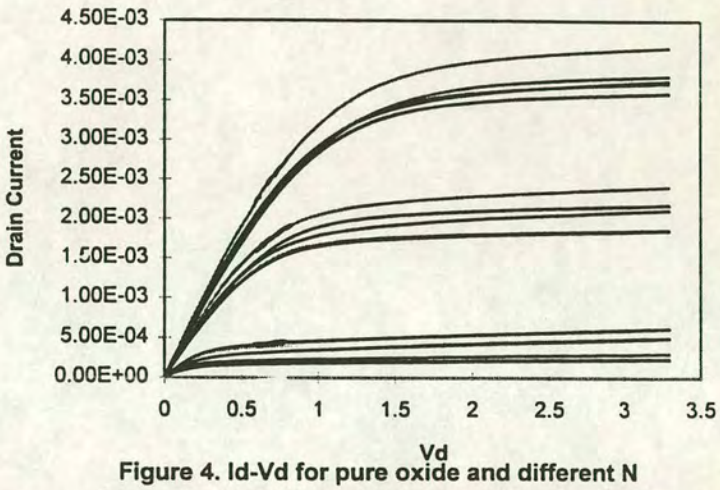
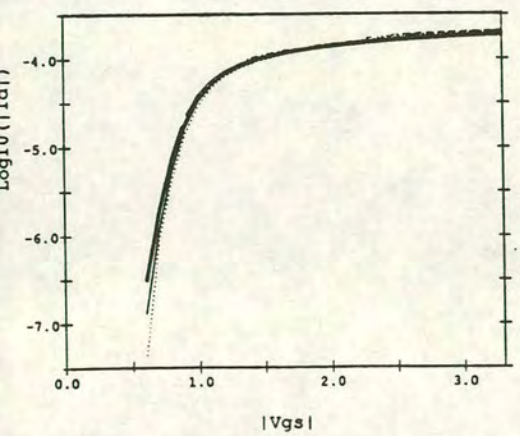


Figure 4. Id-Vd for pure oxide and different N implants. The different Id's for a given Vg result from Tox changes



re 5. Examples of log Id-Vg - pure oxide (double ine), 1E14 (single line) & 1.2E14 (dotted line)

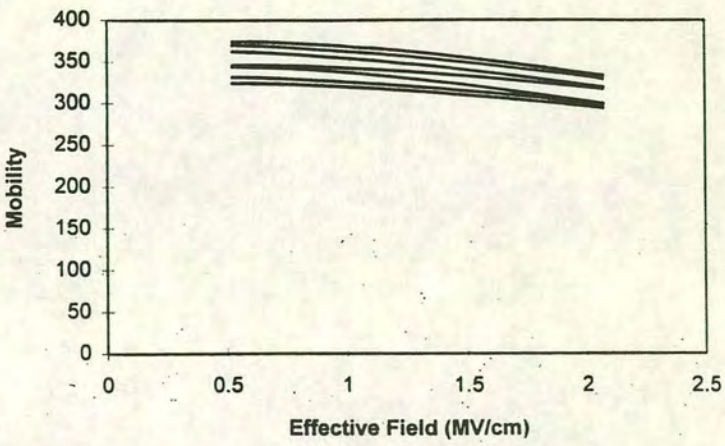


Figure 6 - μ Vs Effective Field for n-MOS transistors, showing that increasing N dose lowers μ (Max N = 2E14)

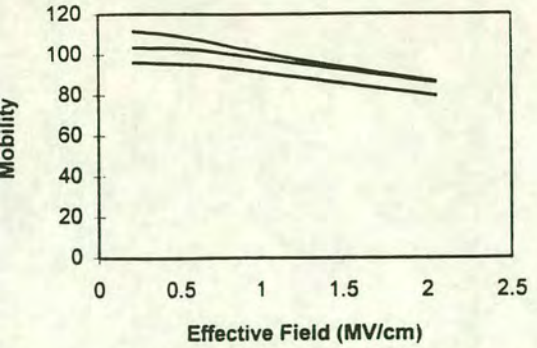


Figure 7. μ Vs Effective Field for p-MOS transistors. Increasing N dose lowers μ

Implant dose (*E14)	RMS roughness (nm)	Correlation length (nm)
pure oxide	.205-.207	16.64-17.16
1.00	.229-.232	13.00-15.41
1.20	.239-.243	15.44-16.01
2.00	.347-.350	17.39-17.71
3.00	.599-.620	19.27-19.87

Figure 8. N-Implant dose vs Surface Roughness & Correlation Length

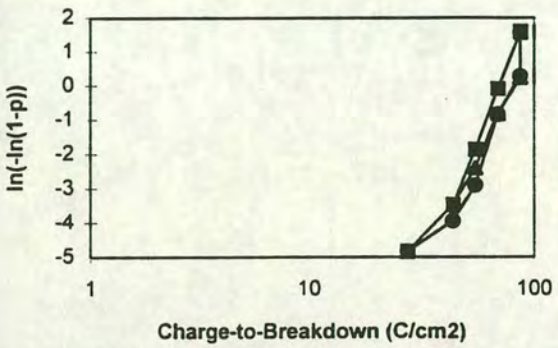


Figure 9. Probability Vs Charge-to-Breakdown for conv. oxides (circle), 1E14 (triangle) & 1.2E14 (square) for p-MOS gates

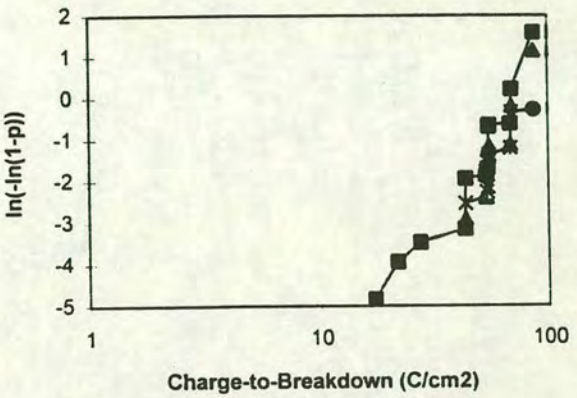


Figure 10. Probability Vs Charge-to-Breakdown for conv. oxides (cirlce), 1E14 (triangle) & 1.2E14 (square) for n-MOS gates

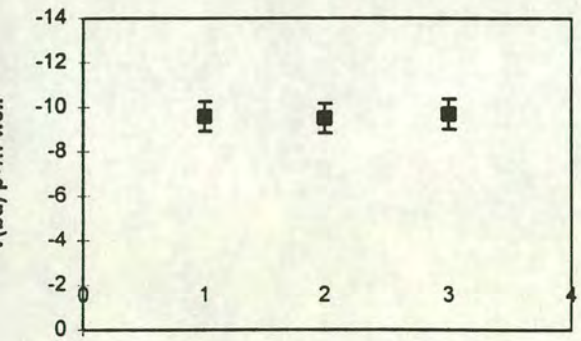


Figure 11. Breakdown Voltage for p+ - n-well junctions for non-implanted (1), 1.0E14 (2) and 1.2 E14 (3). Average of 20 devices

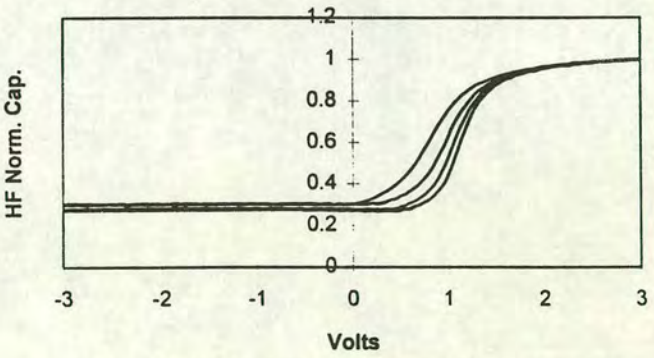


Figure 12. HF Capacitance Vs Voltage. Right to left - oxide (975 C/ 30min), 1E14 (975/30), 1.4E14 (975/30) and oxide (900 C/30 min)

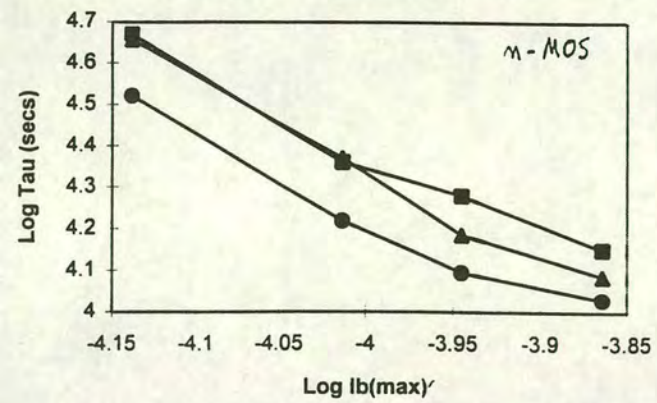


Figure 13. log Ib(max) Vs Iot Tau for pure oxide (circles), 1E14 (triangle) and 1.2E14 (squares)

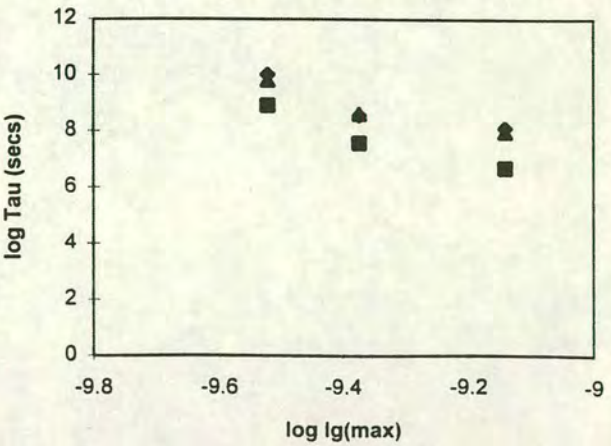


Figure 14. log Ig(max) Vs log Tau for pure oxide (diamonds), 1E14 (triangle) & 1.2E14 (squares) for p-MOS transistors